Word-sense disambiguation in biomedical ontologies

Dissertation

zur Erlangung des akademischen Grades
Doctor rerum naturalium (Dr. rer. nat.)

vorgelegt an der
Technischen Universität Dresden
Fakultät Informatik

von

Dipl.-Bioch. M.Sc. Dimitra G Alexopoulou
geboren am 24 November 1981 in Athen, Griechenland

Gutachter

Prof. Dr.-Ing. Michael Schroeder, Technische Universität Dresden (Betreuender Hochschullehrer)

Prof. Dr. Udo Hahn, Friedrich-Schiller Universität Jena

Tag der Verteidigung: 11 Juni 2010

Dresden, den 8 April 2010

To my parents, George and Maria, and to my grandma, Stefania, who is looking from above

Στους γονεις μου, Γιωργο και Μαρια, και στη γιαγια μου, Στεφανια,
που κοιταει απο ψηλα

Acknowledgments

There are a lot of people who have directly or indirectly helped in making this work possible and I would like to thank.

First of all I would like to thank my supervisor Professor Michael Schroeder for giving me the opportunity to work in such an interdisciplinary and international group, guiding me throughout these years, making things look simple in an inspirational way, caring for us as a group from the scientific side but also from the human side. It has certainly been a great experience for me to work in this group and I will definitely be present in 2017 when we will open the time capsule we all sealed in November of 2007.

I would also like to thank Jörg Hakenberg for an inspiring collaboration on Word Sense Disambiguation, Bill Andreopoulos for pushing for publications from the very first months of working together, Thomas Wächter for sharing his enthusiasm while working at the same office, Heiko Dietze for patiently replying to all database-querying related questions and Andreas Doms for working together in WSD. A big "$\varepsilon\upsilon\chi\alpha\rho\iota\sigma\tau\omega$" goes to George Tsatsaronis for providing valuable and constructive feedback on this thesis.

I would like to especially thank Christof Winter for helping me during my first month in Dresden, when I was still "mute", not speaking a single word of German. He was the person each one of us would like to have around during a fresh start. I can now say "Danke schön" :-).

In my fresh start but also during the years very important was also the contribution of Mandy Gläßer. So, "Vielen vielen Dank" go also to her.

Big thanks also go to the system administrators, Alex Mestiashvili, Nick Dannenberg and Gregor Friedrich, who were at any time available and willing to assist in small and bigger tasks.

I would also like to thank a group of people with which we started as colleagues and we ended up being a lot more, friends, food/beer friends, all-united-against-bad-weather friends, Piled higher and Deeper friends... I could add a lot of tags to them, but not this time. I think 'friends' is enough. So thanks Annalisa Marsico, Anne Tuukkanen, Janine Roy, Conrad Plake, Gihan Dawelbeit, Andreas Henschel.

Special thanks also go to friends with which we have shared the same passion and frustration for science a bit longer than the PhD years, since we started working on Bioinformatics in Athens: Anna Elefsinioti and Evangelia Petsalakis.

More special thanks go the geographically distant but very very very close friends, Rania Limitsiou and Vicky Sagia in Greece and Maria Mirotsou in the US. They know why.

My last but certainly not least thanks go to my parents, George and Maria, and my godmother, Roxanne, for being there anytime and making things sound much easier.

List of Publications

1. Andreopoulos, B., **Alexopoulou, D.**, and Schroeder, M. (2008). **Word sense disambiguation in biomedical ontologies with term co-occurrence analysis and document clustering**. *International Journal of Data Mining and Bioinformatics*, 2(3), 193–215. (Special Issue on Text Mining and Information Retrieval).

2. **Alexopoulou, D.**, Wächter, T., Pickersgill, L., Eyre, C., and Schroeder, M. (2008). **Terminologies for text-mining; an experiment in the lipoprotein metabolism domain**. *BMC Bioinformatics*, 9 Suppl 4, S2.

3. **Alexopoulou, D.**, Andreopoulos, B., Dietze, H., Doms, A., Gandon, F., Hakenberg, J., Khelif, K., Schroeder, M., and Wächter, T. (2009). **Biomedical word sense disambiguation with ontologies and meta-data: automation meets accuracy**. *BMC Bioinformatics*, 10(1), 28.

4. Oliver, H., Diallo, G., de Quincey, E., Kostkova, P., Jawaheer, G., **Alexopoulou, D.**, Habermann, B., Stevens, R., Jupp, S., Khelif, K., Schroeder, M., and Madle, G. (2009). **A user-centred evaluation framework for the Sealife semantic web browsers**. *BMC Bioinformatics*, 10, S14. (Special issue dedicated to the SWAT4LS workshop).

Abstract

With the ever increase in biomedical literature, text-mining has emerged as an important technology to support bio-curation and search. Word sense disambiguation (WSD), the correct identification of terms in text in the light of ambiguity, is an important problem in text-mining. Since the late 1940s many approaches based on supervised (decision trees, naive Bayes, neural networks, support vector machines) and unsupervised machine learning (context-clustering, word-clustering, co-occurrence graphs) have been developed. Knowledge-based methods that make use of the WordNet computational lexicon have also been developed. But only few make use of ontologies, i.e. hierarchical controlled vocabularies, to solve the problem and none exploit inference over ontologies and the use of metadata from publications.

This thesis addresses the WSD problem in biomedical ontologies by suggesting different approaches for word sense disambiguation that use ontologies and metadata. The "Closest Sense" method assumes that the ontology defines multiple senses of the term; it computes the shortest path of co-occurring terms in the document to one of these senses. The "Term Cooc" method defines a log-odds ratio for co-occurring terms including inferred co-occurrences. The "MetaData" approach trains a classifier on metadata; it does not require any ontology, but requires training data, which the other methods do not.

These approaches are compared to each other when applied to a manually curated training corpus of 2600 documents for seven ambiguous terms from the Gene Ontology and MeSH. All approaches over all conditions achieve 80% success rate on average. The MetaData approach performs best with 96%, when trained on high-quality data. Its performance deteriorates as quality of the training data decreases. The Term Cooc approach performs better on Gene Ontology (92% success) than on MeSH (73% success) as MeSH is not a strict is-a/part-of, but rather a loose is-related-to hierarchy. The Closest Sense approach achieves on average 80% success rate.

Furthermore, the thesis showcases applications ranging from ontology design to semantic search where WSD is important.

Contents

List of Figures

List of Tables

Chapter 1

Motivation

1.1 Definition of Open Problems

In the bioinformatics domain, a vast molecular biology literature discusses the relationships between genes, proteins, the biological processes in which they participate, the diseases they are related to, the molecular functions they perform, the cellular location in which they act, and many other types of information. Huge databases with number of records exceeding one million such as Entrez Gene[1] or containing more than 50 million associations such as the UniProt Gene Ontology Annotation[2] (GOA) are under construction tabulating such relationships, but there is a gap between the free text data in articles and the structured data in the databases. Most such databases are manually curated by domain experts and constantly improved in terms of quantity and quality with input from the respective research communities. This process guarantees high data quality and reliability. For instance, annotations of genes and gene products are stored in structured manners (associated functions, phenotypes, etc.), so that they can easily be queried (Camon *et al.*, 2004). Controlled vocabularies and ontologies designed for specific types of annotations reduce the amount of ambiguity for both curation and later access.

Database curators constantly scan the relevant literature to find evidence for new annotations related to their domain. These annotations are standardised terms from controlled vocabularies, often referred to as *ontologies*. For genes and gene products, annotations reflecting functions, locations, and processes are sought (Ashburner *et al.*, 2000). For drugs, it is interesting to find known biochemical pathways and respective (desired and undesired) targets (Degtyarenko *et al.*, 2008; Arikuma *et al.*, 2008; Banville, 2009; Skrzypek *et al.*, 2010). Such facts are often reported in the literature and spread over a large variety of journals and other publication formats.

Databases vs. Literature

Queries across disparate databases are required to exploit available data. However, a lot of data are not yet stored in a structured form. This is due to two main reasons:

1. there is no immediate interest for researchers to submit their findings to (one or more) relevant databases, as scientific publications function as the main instrument for making information accessible and gaining reputation, and

2. the necessary process of manual curation of database entries and annotation needs to maintain a certain quality standard.

Another resource of data are the aforementioned scientific publications themselves. Fairly often, these provide insight into more recent findings than databases. In addition, more information can be found

[1] See http://www.ncbi.nlm.nih.gov/sites/entrez?db=gene
[2] See http://www.ebi.ac.uk/GOA/uniprot_release.html

in texts, such as background knowledge and descriptions of experimental settings, showing broader context as well as in-depth details. Natural language is often more suitable to express facts than the structured form of any database. Moreover, many annotations in databases come in the form of free text, e.g., functions and diseases in UniProt[3]. This shows that scientific publications and other textual descriptions present important resources to be considered when searching for certain information.

Text Mining and Ambiguity

In biomedical text mining, researchers use techniques from *Natural Language Processing*, *Information Retrieval*, and *Machine Learning* to extract desired information from text (Jensen *et al.*, 2006; Winnenburg *et al.*, 2008). Even when the concepts to extract are available in a structured form, such as a controlled vocabulary or ontology, mining them from free text is not always an easy task. For instance, a recent assessment for extracting Gene Ontology terms revealed performances around 20% success rate only (Ehrler *et al.*, 2005). Deciding on the correct sense of the term is also not an easy task; the KDD Cup competition for 2002 (Yeh *et al.*, 2002) challenged researchers with the task of analyzing scientific articles in order to extract information useful for human annotators of the *Drosophila* genome - specifically with identifying all the genes mentioned in an article and determining for each one whether the article reports a relationship between that gene and a gene product (protein and/or mRNA). A complicating factor in the task - and in the biomedical literature in general - is that the term can often refer to the gene, protein or mRNA. Other problems of ambiguity include abbreviations, e.g., whether *MG* refers to *milligram* or *magnesium* (Yu *et al.*, 2002, 2007) and the interpretation of acronyms, e.g., whether or not *COLD* should be interpreted as *chronic obstructive pulmonary disease*.

1.1.1 Open Problem 1: Word Sense Disambiguation in Biomedical Corpora

Terms can have a very specific meaning in biomedical research, but mean other things in other contexts; they can resemble common names, diseases, or common English words. Examples of ambiguous gene names are "Ken and Barbie", "multiple sclerosis" or "the". Some drug names such as "Trial" or "Act" are also ambiguous. In the BioCreAtIvE 2[4] challenge, the results for the gene name normalization task where promising, with up to 86% success rate for human genes (Hakenberg *et al.*, 2007; Wermter *et al.*, 2009). Disambiguation in gene identification can be performed using background knowledge on each gene, such as function, chromosomal location, related diseases, etc (Hakenberg *et al.*, 2008). When extracting ontology terms from text, such valuable information is not present. Hence the two problems are slightly different in that disambiguation in gene identification can use additional data about the entities which the task at hand here cannot.

Identification of ontological terms in literature is in general a challenging problem due to a series of points: *term variation* in natural language text (orthographic, morphological, lexical, structural, acronyms/abbreviations, synonyms, etc.), *synonymity* of ontological terms, *ambiguity*, *stemming* and *missing words*. Coming to ambiguity, ontology term labels can have multiple senses and therefore be ambiguous. The standard problems in ambiguity stem from polysemy and synonymy of words. *Polysemy* means to have multiple meanings; it is an intrinsic property of words (in isolation from text), whereas *ambiguity* is a property of text. Whenever there is uncertainty as to the meaning that a speaker or writer intends, there is ambiguity. So, polysemy indicates only potential ambiguity, and context works to remove ambiguity (Edmonds and Agirre, 2006). The word 'bank', for example, can have several *homonyms/homographs* with clearly different senses including the financial institution, a step or edge as in "snow bank" or "river bank", or other as in the "piggy bank", the 'BANK' gene, protein or mRNA. To demonstrate polysemy, 'bank' as a financial institution can split into the following cloud of related

[3] http://www.uniprot.org/
[4] See http://biocreative.sourceforge.net/biocreative_2.html

senses: the company or institution, the building itself, the counter where money is exchanged, a fund or reserve of money, the funds in a gambling house and the dealer in a gambling house.

The task of disambiguation is hard even for human annotators, when it comes to fine grained sense inventory. According to Halliday and Hasan (1976), the human performance in WSD reaches 97-99% accuracy in coarse grained sense inventories, and 65-70% in fine grained. The average f–measure for inter-annotator agreement for manual GO curation reaches 82% (Camon *et al.*, 2005). With disambiguation being already hard for human annotators, the challenge is even bigger for automated methods which identify ontology terms in text.

Word Sense Disambiguation Research

Word sense disambiguation (WSD) was first formulated as a distinct computational task during the early days of machine translation in the late 1940s, making it one of the oldest problems in computational linguistics. Weaver (1955) introduced the problem as follows:

> "If one examines the words in a book, one at a time through an opaque mask with a hole in it one word wide, then it is obviously impossible to determine, one at a time, the meaning of words. "Fast" may mean "rapid"; or it may mean "motionless"; and there is no way of telling which.
>
> But, if one lengthens the slit in the opaque mask, until one can see not only the central word in question but also N words on either side, then, if N is large enough one can unambiguously decide the meaning."

Weaver acknowledged that context is crucial and recognized the basic statistical character of the problem in proposing that "statistical semantic studies should be undertaken, as a necessary primary step".

During the 1950s, there was a lot of research in estimating the degree of ambiguity in texts and bilingual dictionaries, and applying simple statistical models. Zipf (1949) published the "Law of Meaning" that accounts for the skewed distribution of words by number of senses, meaning that more frequent words have more senses than less frequent words in a power-law relationship. Edmonds (2005) has confirmed this relationship for the British National Corpus. Kaplan (1955) determined that two words of context on either side of an ambiguous word was equivalent to a whole sentence of context in resolving power. Most of the early work set the basis for approaches still followed today. Masterman (1957) used the headings of the categories in Roget's International Thesaurus (Chapman, 1977) to represent the different senses of a word, and then chose the heading whose contained words were most prominent in the context. Madhu and Lytle (1965) calculated sense frequencies of words in different domains – based on their observation that domain constrains sense – and then applied Bayes formula to choose the most probable sense given a context. The problem of WSD was one of the reasons why most of machine translation was abandoned in the 1960s, due to the report from the Automatic Language Processing Advisory Committee (ALPAC report) (ALPAC, 1966).

WSD was revisited in the 1970s within artificial intelligence (AI) research on natural language understanding. Wilks developed "preference semantics", where the system used selectional restrictions and frame-based lexical semantics to find a consistent set of word senses for the words in a sentence (Wilks, 1975). In Hirst's system (Hirst, 1987), a word was gradually disambiguated as information was passed between the various modules (including a lexicon, parser, and semantic interpreter) in a process he called "Polaroid Words". "Proper" knowledge representation was important; knowledge sources had to be hand-crafted.

In the 1980s large-scale *lexical resources* and *corpora* became available and handcrafting could be replaced with knowledge extracted automatically from the resources (Wilks *et al.*, 1990). Lesk (1986) used the overlap of word sense definitions in the Oxford Advanced Learner's Dictionary of Current

English (OALD) to resolve word senses. Given two or more target words in a sentence, the pair of senses whose definitions have the greatest lexical overlap were chosen.

With dictionary-based disambiguation the relationship of WSD to lexicography became explicit. For example, Guthrie *et al.* (1991) used the subject codes (e.g., Economics, Engineering) in the Longman Dictionary of Contemporary English (LDOCE) (Procter, 1978) on top of Lesk's method. Yarowsky (1992) combined the information in Roget's International Thesaurus with co-occurrence data from large corpora in order to learn disambiguation rules for Roget's classes, which could then be applied to words in a manner reminiscent of Masterman (Masterman, 1957). Although dictionary-based methods were useful in some cases such as homographs, they were not robust, since dictionaries lacked complete coverage of information on sense distinctions.

Three key points for disambiguation during the 1990s were the development and publication of WordNet (Fellbaum, 1998), the statistical revolution in Natural Language Processing (NLP), and the first Senseval/Semeval evaluation contest (see next paragraph). WordNet pushed research forward because it was both computationally accessible and hierarchically organized into word senses called *synsets*. Today, English WordNet (together with wordnets for other languages[5]) is the most-used general sense inventory in WSD research (Edmonds and Agirre, 2006).

SemEval

SemEval (former Senseval, http://www.senseval.org/) is an evaluation contest for WSD systems that runs every 3 years since 1997. It organizes and runs evaluation and related activities to test the strengths and weaknesses of WSD systems with respect to different words, different aspects of language, and different languages. SemEval is run by a small committee under the auspices of ACL-SIGLEX (the Special Interest Group on the LEXicon of the Association for Computational Linguistics).

Before SemEval, it was extremely difficult to compare and evaluate different systems because of disparities in test words, annotators, sense inventories, and corpora. Gale *et al.* (1992b) noted that "the literature on word sense disambiguation fails to offer a clear model that we might follow in order to quantify the performance of our disambiguation algorithms", and so they introduced lower bounds (choosing the most frequent sense) and upper bounds (the performance of human annotators). However, these could not be used effectively until sufficiently large test corpora were generated. Semeval was first discussed in 1997 (Resnik and Yarowsky, 1999; Kilgarriff and Palmer, 2000) and has grown into the primary forum for researchers to discuss and advance the field. Its main contribution was to establish a framework for WSD evaluation that includes standardized task descriptions and an evaluation methodology. It has also focused research, enabled scientific rigor, produced benchmarks, and generated substantial resources in many languages (e.g., sense-annotated corpora), thus enabling research in languages other than English.

WSD in the Biomedical Domain

In the biomedical domain, WSD has become a hot topic in the last years. The challenge here is the rapid growth of the biomedical literature in terms of new words and their senses, with the situation getting worse with the use of abbreviations and synonyms. Quoting Ide and Véronis (1998), "WSD work has come full circle, returning most recently to empirical methods and corpus-based analyses that characterized some of the earliest attempts to solve the problem". This illustrates the exact need in the case of the biomedical domain; the development of statistical approaches that utilize "established knowledge" (like thesauri, dictionaries, ontologies and lexical knowledge bases) and require no or only some parsing of the text in order to perform the correct annotation.

Two main decision points for WSD in the biomedical domain are the granularity to which WSD should be performed and the selection of an appropriate corpus for training and evaluation. Concerning

[5]For wordnets around the globe, see http://www.globalwordnet.org/gwa/wordnet_table.htm

granularity, some tasks are easier than others (e.g., distinguishing between 'bank' as a building vs the 'BANK' gene is easier than 'BANK' gene vs the protein). Concerning the biomedical corpora, those are either few or do not apply universally, mainly due to the time-consuming and labor-intensive process of manual or semi-automatic annotation. Examples of biomedical datasets are the NLM WSD test collection (Weeber *et al.*, 2001), Medstract[6] for acronyms and the BioCreAtIvE[7] set for mouse, fruitfly, and yeast. However, depending on the task, researchers need to collect their own gold standard datasets.

WSD approaches can be broadly distinguished as *supervised* and *unsupervised*, with a further distinction between *knowledge-based* (or *knowledge-rich*, or *dictionary-based*, or using *established knowledge*) and *corpus-based* (or *knowledge-poor*) (Schuemie *et al.*, 2005; Edmonds and Agirre, 2006; Navigli, 2009). In the biomedical domain researchers have focused on supervised (Hatzivassiloglou *et al.*, 2001; Liu *et al.*, 2004; Gaudan *et al.*, 2005; Pahikkala *et al.*, 2005) and knowledge-based methods (Schijvenaars *et al.*, 2005; Humphrey *et al.*, 2006; Hakenberg *et al.*, 2008; Farkas, 2008) to perform gene name normalization and resolve abbreviations. According to the BioCreAtIvE 2 challenge, the former problem can be solved with up to 86% success rate for human genes, which are challenging with 1.03 genes per name on average (Hakenberg *et al.*, 2008). A more detailed description of WSD algorithms is given in Chapter 2, Section 2.1.1.

> **Open problem 1:** Word sense disambiguation (WSD) is required for the accurate analysis of text in many applications. Since 2004, the most active domain-specific application area for WSD seems to be bioinformatics (Liu *et al.*, 2004; Schuemie *et al.*, 2005; Edmonds and Agirre, 2006). Classical approaches to WSD use co-occurring words or terms. However, most treat ontologies as simple terminologies, without making use of the ontology structure or the semantic similarity between terms. We explore disambiguation of terms in abstracts of biomedical publications using co-occurrence analysis, document clustering, the ontology structure and semantic similarity between terms, as well as metadata.

1.1.2 Open problem 2: Text mining and WSD in Biomedical Terminologies

As already mentioned, ambiguity is widespread among the biomedical ontologies. 'CAM', for example, can stand for 'constitutively active mutants', 'cell adhesion molecule', or 'complementary alternative medicine'. 'Embryo' can refer to 'human embryo' or 'mouse embryo'; 'male' can refer to a human patient or an animal; 'development' can refer to 'embryo development', 'software development', 'cell culture development', 'staff development', etc. There have been some efforts to use several biomedical ontologies in automated document retrieval and annotation. Here we address the major obstacles faced during the development of a biomedical ontology for use in text-mining.

Text Mining in the Life Sciences

The World Wide Web contains a huge amount of data and information for many topics. However, this information cannot be automatically processed without the presence of semantics associated with it. The *Semantic Web* is a vision of the next generation World Wide Web in which data from multiple sources described with rich semantics are integrated to enable processing by humans as well as software agents. *Text mining* offers methods to automatically extract relevant information contained in free text. In the Semantic Web context, the annotations generated are formalized by ontologies to ensure semantic interoperability between the extracted knowledge embedded in annotations and other knowledge sources. Examples of other knowledge sources are not only other documents (for instance providing summaries or definitions of terms), but also databases, web-services, queryable tools, and comparable dynamic

[6] See http://www.medstract.org/
[7] See http://biocreative.sourceforge.net/resources.html

resources. In the life sciences, such dynamic resources provide data on genes or proteins (including sequences), online access to bioinformatics tools (for instance for similarity searches or multiple sequence alignment), etc. For the purpose of document annotation, most text mining methodologies rely on a dataflow whose core components are the following:

- a *Natural Language Processing (NLP) pipeline* comprises different modules and techniques used to analyze text;
- a *term extractor* module finds all occurrences of an entity in the corpus using information produced by the NLP pipeline and ontology instances;
- a *relation extractor* module is used to extract the relation instances that hold between terms. For this purpose, it can use information embedded in the ontology (the relationship hierarchy) and information produced by the NLP pipeline;
- an *annotation generator* collects information generated by all modules and generates a structured annotation based on the ontology. This annotation can be stored separately or embedded in the text document. In the Semantic Web context, most systems use the RDF language to represent these annotations.

Natural Language Processing (NLP)

Text analysis comprises several distinct stages beginning by breaking the text in words until the presentation of its contents. Natural language processing (NLP) systems implement either the totality of these stages, or a combination of certain stages. The complete text analysis must go through the following steps:

- Morphological analysis: identification of word variations (plural form, abbreviation, etc.) and assigning some lexical information to each word (category, gender, number, etc.);
- Syntactic analysis: identifying the syntactic structures associated to each phrase (subject, verb, object, etc.);
- Semantic analysis: building a set of semantic representations from the syntactic trees;
- Pragmatic analysis: identifying discourse items associated to each text.

These steps need the use of different techniques which include tokenization, PoS tagging and parsing:

- *Tokenization* is the process of breaking the text into its constituent units called tokens. Tokens may vary in granularity depending on application but the most common method of tokenization is the fragmentation of text into words and sentences (sentences splitting).
- The *PoS (part of speech) tagging* is the annotation of words with their appropriate PoS tags taking into account their context within the sentence. The most common tags are: article, noun, verb, adjective, preposition, number, etc. Commonly, the PoS tagging systems are based on rules taggers or on probabilistic models.
- *Parsing* is the process of analyzing an input sequence in order to determine its grammatical structure with respect to a given formal grammar. In NLP, it allows to determine a complete syntactic structure of a sentence. For example, the output of a linguistic role parser is a tree, whose leaves correspond to individual words in the text, and nodes represent linguistic roles, such as Subject, Object, Verb, etc. A particular form of parsing, called *shallow parsing*, consists of computing word sequences (phrases), which are a set of syntactically related words. Each phrase is then tagged by specific predefined tags to annotate noun, verb, and adjective phrases.

Finding Ontology Terms in Text

In biomedical text mining (Jensen *et al.*, 2006), researchers aim to solve problems in automatic annotation of documents by using techniques from natural language processing, information retrieval, and machine learning. But the problems are hard: in a competition to assign Gene Ontology terms to given proteins that occur within given publications, Ehrler et al. achieved the best results with only 20% success rate (Ehrler *et al.*, 2005). The evaluation was based on a manually curated data set of proteins, documents, and Gene Ontology terms called GOA (Camon *et al.*, 2004). The difficulty of automating manual annotation is evident from the fact that only as few as 15% of manually annotated terms appear literally in the associated abstracts. When annotating documents, different sections of the document have varying values. Title and Abstract concisely summarize an article and will therefore have fewer terms than Materials and Methods, which may contain terms not capturing the topic of the document. Ontologies/taxonomies such as the Gene Ontology (GO) (Ashburner *et al.*, 2000) and the Medical Subject Headings[8] (MeSH) (Nelson *et al.*, 2001) have been designed to annotate data and to function as classification schemes, rather than to build novel search engines. Indeed, the identification of ontology terms in text is a difficult problem. Text mining approaches use a variety of techniques such as machine learning, information content of words and alignment techniques (Couto *et al.*, 2005; Doms and Schroeder, 2005; Ehrler *et al.*, 2005) to tackle this task. Typical problems that arise from mining life scientific literature are:

- **Stemming**: often words will appear in different forms, such as 'binding' and 'binds', which can be reduced to their stem 'bind'. However, reducing 'dimerization' to 'dimer' is not valid, since the former describes the process, while the latter the outcome. Reducing 'organization' to 'organ' is clearly not valid, since they belong to a completely different context.

- **Missing words**: the text "...tyrosine phosphorylation of a recently identified STAT family member..." should match the ontology term "tyrosine phosphorylation of STAT protein"; the text "...a transcription factor that binds..." should match the ontology term "transcription factor binding"; the text "alkaline phosphatase" should match the ontology term "alkaline phosphatase activity". In general, matching is allowed to ignore words such as "of", "a", "that", "activity", but obviously not "STAT" or "alkaline phosphatase".

- **Format of terms**: ontology terms may contain commas, hyphens, brackets, etc. which require special treatment. For 'thioredoxin-disulfide' the dash can be dropped, for 'hydrolase activity, acting on ester bonds' the clause after the comma is important, but unlikely to appear as such in text. Terms containing additions such as '(sensu Insecta)' contain important contextual information, but are also unlikely to appear in text.

- **Synonymity**: ontology terms might not appear literally in a text, but authors use synonyms for the same concept instead. When searching for 'digestive vacuole', a user would want to find references that use 'phagolysosome'; mentioning of a 'ligand' refers to the concept of 'binding'; an 'entry into host' might appear as an 'invasion of host'.

- **Word sense disambiguation (WSD)**: Terms can have a very specific meaning in biomedical research, but mean other things in other contexts. Examples are cell, development, envelope, spindle, death, growth, regeneration, transport, host, reproduction, circulation, and many others. Protein names such as 'Ken and Barbie', 'multiple sclerosis' or 'the' that resemble common names, diseases, or common English words are especially hard to disambiguate. The same problems arise from drug names like 'Trial' or 'Act'.

[8] See http://www.nlm.nih.gov/mesh/meshhome.html

In contrast to ontologies designed primarily for annotating biological objects, there is a clear distinction to ontologies designed for text mining. As far as the biomedical ontologies are concerned, during the last years there have been major efforts in the biological community for organizing biological concepts in the form of controlled terminologies or ontologies (Eilbeck *et al.*, 2005; Whetzel *et al.*, 2006; Ashburner *et al.*, 2000; Evsikov *et al.*, 2004). A key difference between terminologies and ontologies is that the former lack the semantic depth of the latter. However, when it comes to design, terminologies can serve as basis for ontologies and vice-versa. An example where a terminology can serve for ontology is that of the Gene Ontology (Ashburner *et al.*, 2000), which provides a controlled vocabulary to describe gene and gene products in any organism. On the other side, the Gene Ontology Next Generation (GONG) project (Wroe *et al.*, 2003) aims at the migration of current bio-ontologies to a richer and more rigorous status, using formal representation languages like OWL. Examples of true ontologies are the GALEN project (Rector *et al.*, 1996) and the Systematized Nomenclature of Medicine (SNOMED) (Spackman, 2004) which are based on Description Logic for concept representation and the Foundational Model of Anatomy (FMA) (Rosse and Mejino, 2003) which is based on frames representing information about anatomical classes, designed so that content can be maintained as a dynamic resource and can be used as terminologies. The OBO Relation Ontology (Smith *et al.*, 2005) has been designed to promote interoperability of ontologies and support new types of automated reasoning about the spatial and temporal dimensions of biological and medical phenomena. The Relation Ontology assists the ontology development process by providing consistent and unambiguous formal definitions of the relational expressions used in biomedical ontologies. In this manner, developers and users avoid errors in docing and annotation.

Semantic meta-information provided in the form of ontologies has proven useful in order to search (Doms and Schroeder, 2005) or index large collections of documents (e.g., MeSH for indexing MEDLINE (Nelson *et al.*, 2001)). Meta-information found for text documents is often general (keyword list) or still too complex for an automated evaluation (article abstract). Finding terms of controlled vocabularies in text overcomes this shortage, while relations between terms provide the necessary navigation structures.

Ontological background knowledge can serve to answer questions with knowledge-based search engines, by easing the task of finding relevant documents through the term automatic annotation (Doms and Schroeder, 2005; Mueller *et al.*, 2004; Perez-Iratxeta *et al.*, 2003; Wermter *et al.*, 2009; DeLuca *et al.*, 2009). In the domain of lipoprotein metabolism, for example, in case of a syndrome, such as the "metabolic syndrome", in a properly designed ontology the articles retrieved will contain symptoms and other characteristics for it (e.g., type II diabetes, hypertension, insulin resistant, low HDL, hypertension, all of them being parts of the metabolic syndrome).

Open problem 2: Which are the common obstacles during the design of an ontology to be used for text mining? Can automatic term recognition (ATR) methods assist the ontology generation process?

In Chapter 4 we share the experience acquired during the manual development of a lipoprotein metabolism ontology (LMO) to be used for text-mining. We provide guidelines for the design of this ontology and describe the common obstacles during the process. We compare the manually created ontology terms with the automatically derived terminology from four different automatic term recognition (ATR) methods.

1.1.3 Overview

In the following chapters we address the aforementioned open problems, starting with an overview of the existing research in each field (see Figure 1.1 for an overview).

Chapter 2 provides an overview of research in word sense disambiguation in general and more specific in the biomedical domain and continues with the use of ontologies in the life sciences and in text mining.

Chapter 3 addresses open problem 1 by suggesting different approaches for word sense disambiguation that use ontologies and metadata. These approaches are compared to each other when applied to benchmark datasets especially collected for the purpose of disambiguation in the biomedical domain. The question whether ontologies can play an important role to improve disambiguation is investigated.

Chapter 4 addresses open problem 2 by suggesting guidelines for the design of terminologies/ontologies for use in text mining, based on the experience acquired during the manual development of a lipoprotein metabolism ontology (LMO) to be used for text-mining by researchers at Unilever. A comparison between the manually created ontology terms with the automatically derived terminology from four different automatic term recognition (ATR) methods is also performed to investigate how automatic methods can help decrease development time and provide support for the identification of domain-specific vocabulary.

Chapter 5 shows use cases of word sense disambiguation in ontology-based text-mining and more specifically in biomedical document retrieval with the GoPubMed semantic search engine (Doms and Schroeder, 2005) and in mouse-anatomy-specific document retrieval (with the MousePubMed variant of GoPubMed). It also describes a user-centred evaluation framework developed to evaluate Semantic Web Browsers, showing the readiness of common users to exploit the benefits of the semantic web in the life sciences domain.

Figure 1.1 provides an overview of the structure of the current document.

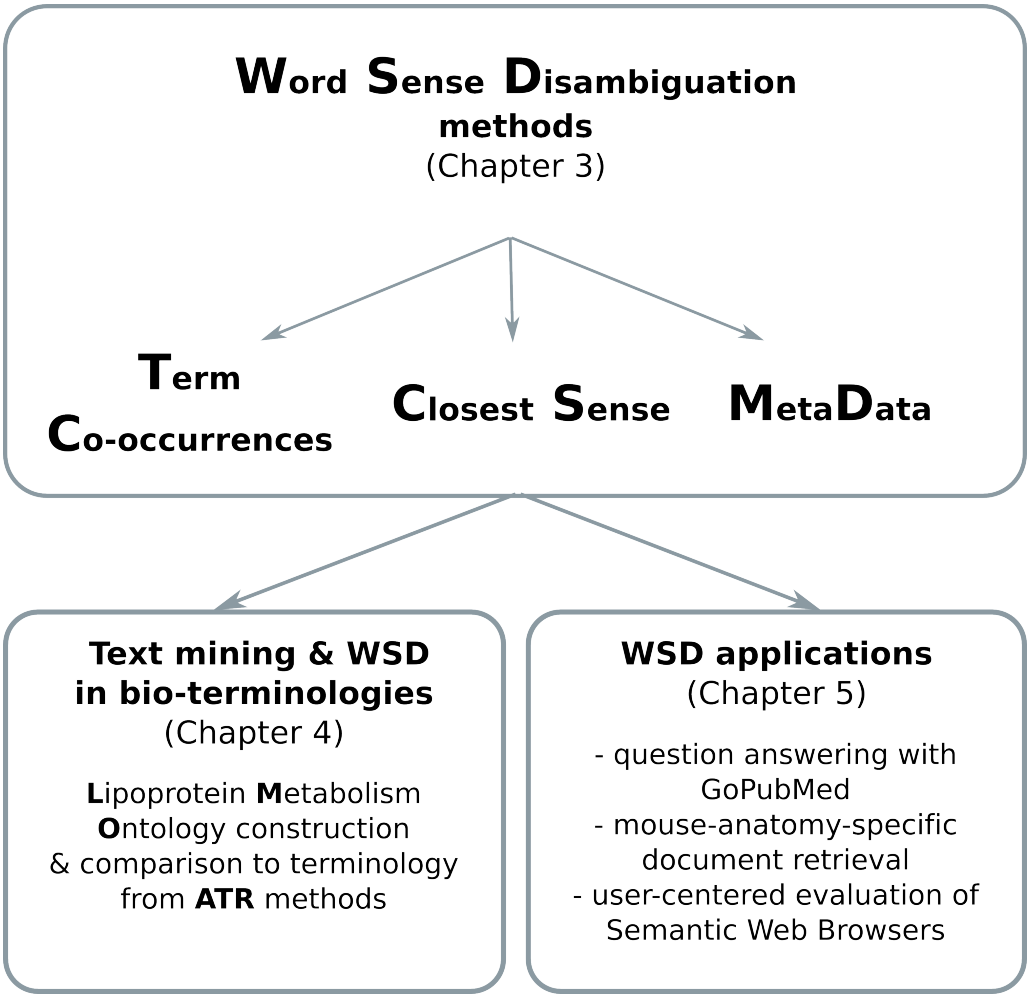

Fig. 1.1: Thesis overview. Chapter 3 suggests different approaches for word sense disambiguation that use ontologies and metadata. Chapter 4 describes problems during the construction of a Lipoprotein Metabolism Ontology and includes a comparison of the manually created terminology to terminologies automatically created by four different automatic term recognition (ATR) methods. Chapter 5 shows applications of word sense disambiguation in ontology-based text-mining and more specifically in question answering with GoPubMed and in mouse-anatomy-specific document retrieval (MousePubMed). It also describes a user-centred evaluation framework developed to evaluate Semantic Web Browsers, showing the readiness of common users to exploit the benefits of the semantic web in the life sciences domain.

Chapter 2

Introduction

2.1 Word Sense Disambiguation (WSD)

Since the announcement of the Human Genome in 2000, some 200 model organisms have been sequenced and novel sequencing technologies will lead to a further increase in data generation. A prime task after sequencing a genome is the identification of genes and their annotation with relevant functions, processes, and cellular components. In order to facilitate the comparison of genomes, biologists devised a shared, species independent vocabulary, the Gene Ontology (GO) (Ashburner *et al.*, 2000), with some 20,000 terms and synonyms. Annotation of novel genomes with the Gene Ontology is a manual process, in which editors read relevant literature for a gene and then decide on suitable annotation. However, manual annotation is labor-intensive: currently, there are several million genes and proteins of almost 60,000 different species represented in the public databases, but only approximately 500 of these species have had GO terms manually assigned in GOA, the Gene Ontology Annotation (Camon *et al.*, 2005, 2004). Presently, much effort is devoted to automating or aiding the annotation process (Jensen *et al.*, 2006). In a recent text mining competition, BioCreative[1], one task consisted in the identification of suitable Gene Ontology terms for a given gene and document. As reported by Ehlrer (Ehrler *et al.*, 2005), the best result in this category achieved only 20% accuracy. Identification of Gene Ontology terms in literature is in general a challenging problem (Doms and Schroeder, 2005). As already mentioned in Section 1.1.2, typical problems that arise from mining life scientific literature are *stemming*, *missing words*, *format of terms* and *ambiguous terms*. The last problem is particularly challenging and various approaches ranging from the use of tagged corpora, dictionaries and thesauri to supervised and unsupervised machine learning have been tried (Xu *et al.*, 2006; Schuemie *et al.*, 2005; Navigli and Velardi, 2005; Liu *et al.*, 2002; Navigli *et al.*, 2003; Schijvenaars *et al.*, 2005; Pahikkala *et al.*, 2005; Gaudan *et al.*, 2005).

Coming to the example of the ambiguous term development, the Gene Ontology defines 'development' as follows:

> The biological process whose specific outcome is the progression of an organism over time from an initial condition (e.g., a zygote, a young adult or a young single celled organism) to a later condition (e.g., a multicellular animal, an aged adult or a mature single celled organism).

It should be noted here that we refer as *True* or *True Positive* to the sense of the ambiguous term that corresponds to the one included in the ontology (biological development in GO, psychological inhibition in MeSH, mitotic spindle in GO, etc.) and as *False* to all other senses (e.g., in the context of software/algorithm development, staff development, Ministry of Development, developing country, method development, etc.)

[1] See http://www.mitre.org/public/biocreative/

Title: Genetic imprinting and embryonic **development**
Authors: Yin LJ, Huang HF.
Journal: Zhejiang Da Xue Xue Bao Yi Xue Ban. 2007 Sep;36(5):509-14.

Abstract: **Erasure, establishment and maintenance of genetic imprinting are indispensable for normal embryonic *development***. All these processes depend on accurate expression and intimate cooperation of kinds of DNA methyltransferases. **Many genetic syndromes and embryo *developmental* anomalies are caused by abnormality of genetic imprinting. Genetic imprinting is important for the nucleus totipotential of primordial germ cell, maturation of gamete,growth and *development* of embryo, structure and function of placenta as well as postnatal growth and *development* of individuals**.

PMID: 17924473 [PubMed - indexed for MEDLINE]

Fig. 2.1: Example of the *true* sense for the ambiguous term 'development' in the abstract of an article in PubMed. The true sense corresponds to the sense of development in the available terminology. Here the terminology is the Gene Ontology and the GO sense is that of *embryonic development*.

Some examples of text in PubMed abstracts containing the *True* sense for development are the following:

- "Erasure, establishment and maintenance of genetic imprinting are indispensable for normal embryonic development.", see also Figure 2.1.
- "Arabidopsis ribonucleotide reductases are critical for cell cycle progression, DNA damage repair, and plant development."
- "Homeodomain-containing proteins are transcription factors that regulate the coordinated expression of multiple genes involved in development, differentiation and malignant transformation."
- "Involvement of the TRAP220 component of the TRAP/SMCC co-activator complex in embryonic development thyroid hormone action."
- "Lymphedema-distichiasis (LD) is an autosomal dominant disorder that classically presents itself as lymphedema of the limbs, with variable age at onset, and double rows of eyelashes (distichiasis). Other complications may include cardiac defects, cleft palate, extradural cysts and photophobia, suggesting a defect in a gene with pleiotrophic effects acting during development."

Some examples of text in PubMed abstracts containing a *False* (or *False Positive* for automatic annotation by GoPubMed) sense for development are the following (see also Figures 2.2, 2.3, 2.4):

- "The development of the Na+ gradient during illumination thus, plays an important role in energy coupling", see also Figure 2.2.
- "The recent discovery of several hypothalamic factors involved in the regulation of anterior pituitary function and the development of sensitive immunocytochemical techniques have greatly contributed to..."
- "Limited diagnostic and therapeutic interventions should be addressed as separate entities in the development of the patient care plan."
- "Academic research, especially university research, tends to be substituted by development innovation for the production process."

Title: Existence of electrogenic hydrogen ion/sodium ion antiport in Halobacterium halobium cell envelope vesicles.
Authors: Lanyi JK, MacDonald RE.
Journal: Biochemistry. 1976 Oct 19;15(21):4608-14.

Abstract: ...Glutamate transport appears to be energized only by the Na+ gradient. **The *development* of the Na+ gradient during illumination thus plays an important role in energy coupling**. The results obtained are consistent with the existence of an electrogenic H+/Na+ antiport mechanism (H+/Na+ greater than 1) in H halobium which facilitates the uphill Na+ efflux...

PMID: 9978 [PubMed - indexed for MEDLINE]

Fig. 2.2: Example of a *false* sense for the ambiguous term 'development' in the abstract of an article in PubMed. To a human reader, it is clear that development here does not correspond to biological development (as in the GO sense). However, this is not easy to automatically extract; the context remains of biomedical nature.

Title: Immunohistochemical analysis of accelerated graft atherosclerosis in cardiac transplantation.
Authors: Louie HW, Pang M, Lewis W, Drinkwater DC, Laks H.
Journal: Curr Surg. 1989 Nov-Dec;46(6):479-83.

Abstract: HHT was performed between minimally genetic mismatched inbred strains of rats. There was no evidence of rejection and immunosuppressive therapy was not instituted. Immunohistochemical analysis using peroxidase conjugated monoclonal anti-rat ASMA of cardiac arterioles in which AGAS developed revealed a decreased peroxidase signal. **The data suggest that modulation of actin expression in subintimal cells of cardiac arterioles may play a critical role in the pathologic *development* of AGAS.**

PMID: 2620541 [PubMed - indexed for MEDLINE]

Fig. 2.3: Example of a *false* sense for the ambiguous term 'development' in the abstract of an article in PubMed. Development here is falsely annotated as 'heart development'.

Title: Corticosteroid cataracts following kidney transplantation. Investigations on the influence of additional factors upon the development of opacities
Authors: Koch HR, Weikenmeier P, Siedek M.
Journal: Albrecht Von Graefes Arch Klin Exp Ophthalmol. 1975;194(1):39-53.

Abstract: 15 patients with kidney transplants were followed up for a period of 3 to 45 months. **All but one *developed* lenticular opacities in the posterior subsapsular region.** The opacities were classified according to their severity. It could be shown that the cataract index was correlated to the total amount of corticosteroids given. There is probably an additional effect of age and azathioprin therapy. **Possibly, the time of treatment with intermittent hemodialysis also influences cataract *development*.**

PMID: 1092198 [PubMed - indexed for MEDLINE]

Fig. 2.4: Example of a *false* sense for the ambiguous term 'development' in the abstract of an article in PubMed. "Development of opacities" or "cataract development" here are falsely annotated as the biological sense.

Title: Monocytes and histiocytes in cell cultures of cerebrospinal fluid. Morphology of cultured CSF cells.
Author: Dommasch D.
Journal: J Neurol. 1975 Jun 9;209(2):103-14.

Abstract: A method of CSF cell culturing, based on observations of cultured cells isolated from 700 CSF specimens obtained for routine diagnostic procedures by lumbar puncture from patients who had no proven or suspected neoplastic disease, is described which enables the demonstration of proliferating mononuclear elements even when they are present in specimens with low cell count. **Spread on surfaces of plastic and glass material, monocytes and histiocytes in CSF cell cultures can appear as polygonal or crescent shaped epitheloid cells, may assume *spindle* shapes, or transform into multinucleated giant cells.** Some cells given rise to clones with different rates of proliferation, up to the formation of a monolayer....

PMID: 51047 [PubMed - indexed for MEDLINE]

Fig. 2.5: Example of a *false* sense for the ambiguous term 'spindle' in the abstract of an article in PubMed. Spindle in "spindle shapes" here is falsely annotated as 'mitotic spindle'.

The difference of the former two examples from the latter two is that the former will contain other Gene Ontology terminology, while the latter are about general topics and thus, they do not contain any other GO terms. The context of the former remains biomedical, while the context of the latter will be completely different, making the disambiguation easier.

There also exist cases where documents do not contain the term 'development' literally, but they are on developmental biology according to the curators (e.g., in GOA). We call these False Negatives (FNs) because if the term does not literally appear, the document will not be automatically annotated with this term. Examples of such documents are:

- "RYBP, a new repressor protein that interacts with components of the mammalian Polycomb complex, and with the transcription factor YY1." The protein YY1 is annotated in the Uniprot sequence database as 'development'.
- "Virtual cloning and physical mapping of a human T-box gene, TBX4". The protein TBX4 is annotated in the Uniprot sequence database as 'development'.
- "Isolation of two novel WNT genes, WNT14 and WNT15, one of which (WNT15) is closely linked to WNT3 on human chromosome 17q21". The protein WNT14 is annotated in the Uniprot sequence database as 'development'.
- "EDF-1, a novel gene product down-regulated in human endothelial cell differentiation". The protein EDF-1 (Endothelial differentiation-related factor 1) is annotated in the Uniprot sequence database as 'development'.

Word Sense Disambiguation (WSD) deals with relating the occurrence of a word in a text to a specific meaning, which is distinguishable from other meanings that can potentially be related to that same word (Schuemie *et al.*, 2005). WSD is essentially a classification problem: given an input text and a set of sense tags for the ambiguous words in the text, assign the correct senses to these words. Sense assignment often involves two assumptions: *a.* within a discourse, e.g., a document, a word is only used in one sense (Gale *et al.*, 1992b) and *b.* words have a tendency to exhibit only one sense in a given collocation - neighbouring words (Yarowsky, 1993).

2.1.1 Algorithms for Word Sense Disambiguation

As already mentioned in Section 1.1.1, WSD approaches can be broadly distinguished as *supervised* and *unsupervised*, with a further distinction between *knowledge-based* (or *knowledge-rich*, or *dictionary-based*, or using *established knowledge*) and *corpus-based* (or *knowledge-poor*) (Schuemie *et al.*, 2005; Edmonds and Agirre, 2006; Navigli, 2009):

- *Supervised WSD* approaches use machine-learning techniques to learn a classifier from labeled training sets, that is, sets of examples encoded in terms of a number of features together with their appropriate sense label (or class).
- *Unsupervised WSD* methods are based on unlabeled corpora, and do not exploit any manually sense-tagged corpus to provide a sense choice for a word in context.
- *Knowledge-based WSD* approaches rely on the use of external lexical resources, such as machine-readable dictionaries, thesauri, ontologies, etc.
- *Corpus-based WSD* approaches do not make any use of the aforementioned knowledge resources for disambiguation. They are based on information from corpora that can be either 'raw' (without any annotation for word senses), 'sense-tagged' or 'automatically-tagged' (for details on corpora, see Unstructured Resources below).

In Figure 2.6, the WSD approaches are put on a bidimensional space, according to the *degree of supervision* and the *amount of knowledge used* (Navigli, 2009). The degree of supervision is expressed by the ratio of sense-annotated data to unlabeled data used by the system: a system is defined as *fully* (or *strongly*) *supervised* if it exclusively employs sense-labeled training data, *semisupervised* and *weakly* (or *minimally*) *supervised* if both sense-labeled and unlabeled data are employed in different proportions to learn a classifier, *fully unsupervised* if only unlabeled plain data is employed. The amount of knowledge concerns all other data employed by the system, including dictionary definitions, lexico-semantic relations, domain labels, and so on. It is hard to quantify the degree of supervision and the amount of knowledge as discrete numbers, so putting on the plane specific methods discussed next is not feasible. However, the letters (a) to (i) in Figure 2.6 give the approximate position of general approaches:

(a) fully unsupervised methods, which do not use any amount of knowledge (not even sense inventories)

(b) minimally supervised approaches, requiring a minimal amount of supervision

(c) semi-supervised approaches, requiring a partial amount of supervision

(d) supervised approaches (machine-learning classifiers).

Associating other points in space with specific approaches is more difficult and depends on the specific implementation of each method. However, most knowledge-based approaches relying on structural properties (g), such as the graph structure of semantic networks, use more supervision and knowledge than those based on gloss overlap (e) or methods for determining word sense dominance (f). Finally, domain-driven approaches, which often exploit hand-coded domain labels, can be placed around point (h) if they include supervised components for estimating sense probabilities, else around point (i).

Knowledge Sources

Knowledge sources provide data which are essential to associate senses with words. They can vary from corpora of texts, either unlabeled or annotated with word senses, to machine-readable dictionaries, thesauri, glossaries, ontologies, etc. Knowledge sources can be categorized into structured and unstructured ones, as follows (Navigli, 2009):

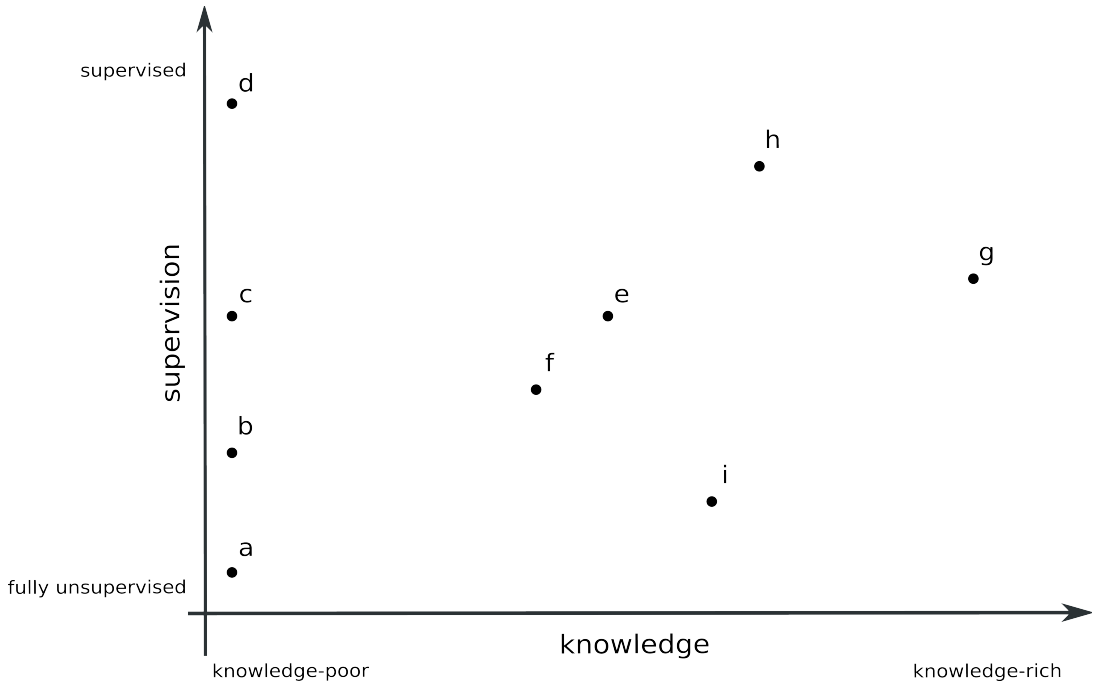

Fig. 2.6: A space of WSD approaches according to the degree of supervision and the amount of knowledge used. At point (a) lie the fully unsupervised methods, which do not use any amount of knowledge to perform the disambiguation. In (b) are the minimally supervised approaches, (c) the semi-supervised approaches and (d) the supervised approaches. The rest of the areas express methods that are combinations of the four categories. These are knowledge-based approaches relying on gloss overlap (e), methods for determining word sense dominance (f), methods based on structural properties (g), and domain-driven approaches which exploit hand-coded domain labels, including supervised components for estimating sense probabilities (h) or not (i). *Adapted from Navigli (2009).*

Structured Resources

- *Thesauri* provide information about relationships between words, like synonymy (e.g., car is a synonym of motorcar), antonymy (representing opposite meanings, e.g., ugly is an antonym of beautiful) and, possibly, further relations. The most widely used thesaurus in the field of WSD is Roget's International Thesaurus (Chapman, 1977).

- *Machine-readable dictionaries* (MRDs) have become a popular source of knowledge for Natural Language Processing since the 1980s, when the first dictionaries were made available in electronic format. Some of these are the Collins English Dictionary, the Oxford Advanced Learner's Dictionary of Current English, the Oxford Dictionary of English, and the Longman Dictionary of Contemporary English (LDOCE) (Procter, 1978). The latter has been one of the most widely used machine-readable dictionaries within the NLP research community, before the diffusion of WordNet[2] (Fellbaum, 1998), presently the most utilized resource for word sense disambiguation in English. WordNet is often considered one step beyond common MRDs, as it encodes a rich semantic network of concepts. For this reason it is usually defined as a computational lexicon.

- *Ontologies* are specifications of conceptualizations of specific domains of interest (Gruber, 1993), usually including a taxonomy and a set of semantic relations. In this respect, WordNet and its extensions can be considered as ontologies. Efforts in a domain-oriented direction include the Open Biomedical Ontologies (OBO) foundry[3], the NCBO BioPortal[4] and the Unified Medical Language System (UMLS) (Bodenreider, 2004), which includes a semantic network providing a categorization of medical concepts. For more details, see Section 2.2.1.

Unstructured Resources

- *Corpora* are collections of texts used for learning language models. Corpora can be sense-annotated or raw (i.e., unlabeled). Both kinds of resources are used in WSD, and are most useful in supervised and unsupervised approaches, respectively. Popular examples of *raw corpora* are the Brown Corpus (Kučera and Francis, 1967) – a million word balanced collection of texts published in the United States in 1967 –, the British National Corpus (BNC) (Clear, 1993) – a 100 million word collection of written and spoken samples of the English language (often used to collect word frequencies and identify grammatical relations between words) –, and the Wall Street Journal (WSJ) corpus (Charniak *et al.*, 2000) – a collection of approximately 30 million words from WSJ. SemCor (Miller *et al.*, 1993) is the largest and most used *sense-annotated corpus*, which includes 352 texts tagged with around 234,000 sense annotations. Other examples of sense-annotated corpora are MultiSemCor (Pianta *et al.*, 2002), the Interest corpus (Bruce and Wiebe, 1994), the Line-Hard-Serve corpora[5] with 4,000 sense-annotated examples for each of the words 'line', 'hard' and 'serve' and the SensEval/SemEval[6] data sets, semantically-annotated corpora from the four evaluation contests. Most of these corpora are annotated with different versions of the WordNet sense inventory. The sense-annotation of corpora has a high degree of variability, mainly due to the variability of domains of interest. Tomanek *et al.* (2007) have used active learning for text corpus annotation, reporting reduction rates for annotation efforts ranging up to 72%.

- *Collocation resources* register the tendency for words to occur regularly with others. Examples include the Word Sketch Engine[7], JustTheWord[8], The British National Corpus collocations[9], the

[2]See http://wordnet.princeton.edu/
[3]See OBO foundry http://www.obofoundry.org/
[4]See BioPortal http://bioportal.bioontology.org/
[5]See http://www.d.umn.edu/ tpederse/data.html
[6]See http://www.senseval.org/
[7]See http://www.sketchengine.co.uk
[8]See http://193.133.140.102/JustTheWord
[9]Available through the SARA system from http://www.natcorp.ox.ac.uk

Collins Cobuild Corpus Concordance[10], etc. Recently, the Google Web1T corpus[11] (Brants and Franz, 2006) has been released; it is a large dataset of text co-occurrences which has rapidly gained large popularity in the WSD community. The Google Web1T corpus contains frequencies for sequences of up to five words in a one trillion word corpus derived from the Web.

- Other resources include *word frequency lists*, *stoplists* (i.e., lists of undiscriminating non-content words, like a, an, the, and so on), *domain labels* (Magnini and Cavaglià, 2000), etc.

As aforementioned, WSD algorithms can be distinguished between *supervised* and *unsupervised*, and *knowledge-based* and *corpus-based* (Schuemie *et al.*, 2005; Edmonds and Agirre, 2006; Navigli, 2009) (see also Table 2.1).

Supervised Machine Learning WSD Approaches

Supervised WSD uses machine-learning techniques to learn a classifier from labeled training sets. The training set used to learn the classifier typically contains a set of examples in which a given target word is manually tagged with a sense from the sense inventory of a reference dictionary. In general, supervised approaches to WSD have obtained better results than unsupervised methods (described later). The supervised WSD approaches can be listed as follows (for a more detailed description of each, see Navigli (2009)):

Decision Lists: A decision list (Rivest, 1987) is an ordered set of rules for categorizing test instances (in the case of WSD, for assigning the appropriate sense to a target word). It can be seen as a list of weighted 'if-then-else' rules. A training set is used for inducing a set of features. As a result, rules of the kind (feature-value, sense, score) are created. The ordering of these rules, based on their decreasing score, constitutes the decision list. Given a word occurrence w and its representation as a feature vector, the decision list is checked, and the feature with highest score that matches the input vector selects the word sense to be assigned. Decision lists have been the most successful technique in the first Senseval evaluation competitions (Yarowsky, 2000). Agirre and Martinez (2000) applied them in an attempt to relieve the knowledge acquisition bottleneck caused by the lack of manually tagged corpora.

Decision Trees: A decision tree is a predictive model used to represent classification rules with a tree structure that recursively partitions the training data set. Each internal node of a decision tree represents a test on a feature value, and each branch represents an outcome of the test. A prediction is made when a terminal node (i.e., a leaf) is reached. A popular algorithm for learning decision trees is the C4.5 algorithm (Quinlan, 1993). In a comparative experiment with several machine learning algorithms for WSD, Mooney (1996) concluded that decision trees obtained with the C4.5 algorithm are outperformed by other supervised approaches suffering mainly from data sparseness due to features with a large number of values and unreliability of predictions due to small training sets.

Naive Bayes: A Naive Bayes classifier is a simple probabilistic classifier based on the application of Bayes' theorem. It relies on the calculation of the conditional probability of each sense S_i of a word w given the features f_j in the context. Naive Bayes compares well with other supervised methods (Mooney, 1996; Ng, 1997; Leacock *et al.*, 1998; Pedersen, 1998; Bruce and Wiebe, 1999).

Neural Networks: A neural network (McCulloch and Pitts, 1943) is an interconnected group of artificial neurons that uses a computational model for processing data based on a connectionist approach.

[10] See http://www.collins.co.uk/Corpus/CorpusSearch.aspx
[11] See http://www.ldc.upenn.edu/Catalog/CatalogEntry.jsp?catalogId=LDC2006T13

	Publication	Data	Backgr. knowledge	Approach	Experiment	Accuracy
Knowledge–based	Schijvenaars *et al.* (2005)	gene definition & abstract vector	5 human gen. dbs & MeSH	cosine similarity	52,529 Medline abstracts 690 human gene symbols	92.7%
	Humphrey *et al.* (2006)	free text	UMLS, Journal Descriptors	Journal Descriptor Indexing (JDI)	45 ambiguous UMLS terms (NLM WSD Collection)	78.7%
	Hakenberg *et al.* (2008)	Medline abstracts	BioCreative-2 GN lexicon & text, EntrezGene, UniProt, GOA	motifs from multiple sequence alignments	BioCreative-2 GN challenge	86%
	Farkas (2008)	Medline abstracts	list of gene senses, EntrezGene	inverse co-author graph	BioCreative GN challenge	97%P
Supervised	Hatzivassiloglou *et al.* (2001)	XML tagged abstracts positional info, PoS	–	naive Bayes, decision trees, inductive rule training	protein/gene/mRNA assignment: 9 million words (mol. biol. journals)	85%
	Ginter *et al.* (2004)	text	–	word count, word cooc	–	86.5%
	Liu *et al.* (2002, 2004)	Medline abstracts	UMLS terms	UMLS term cooc	35 biomedical abbreviations	93%P
	Gaudan *et al.* (2005)	abbreviations in Medline abstracts	–	SVM	build dictionary, use for abbreviations occurring with their long forms	98.5%
	Pahikkala *et al.* (2005)	gene symbol context (n words +/-)	–	SVM	–	85%
Unsupervised	Schütze and Pedersen (1995); Schütze (1998)	document	–	LSA/LSI, 2^{nd} order cooc	170,000 documents, 1013 terms (TREC-1) (Wall Street Journal)	↑ 7-14%
	Pedersen and Bruce (1997)	word cooc, PoS tags	WordNet	average link clustering	13 words, ACL/DCI	73.4%
	Pedersen and Bruce (1998)				Wall Street Journal Corpus	
	Purandare and Pedersen (2004)	–	–	1^{st},2^{nd} order context vectors (coocs within 5 positions)	24 Senseval-2 words, *Line*, *Hard*, *Serve* corpora	44%
	Yarowsky (1995)	text	few tagged data, WordNet	co-training, collocations	12 common Engl. words x 4000 instances	96.5%
	Mihalcea (2004)	–	–	co-training & majority voting	Senseval-2 generic English	↑ 9.8%
	Dorow and Widdows (2003)	–	WordNet	noun coocs, Markov clustering	–	–

Tab. 2.1: Algorithms for Word Sense Disambiguation. In the biomedical domain researchers have focused on knowledge-based (Schijvenaars *et al.*, 2005; Humphrey *et al.*, 2006; Hakenberg *et al.*, 2008; Farkas, 2008) and supervised methods (Hatzivassiloglou *et al.*, 2001; Liu *et al.*, 2004; Gaudan *et al.*, 2005; Pahikkala *et al.*, 2005) to perform disambiguation.

Pairs of *input feature–desired response* are input to the learning program. Cottrell (1989) employed neural networks to represent words as nodes: the words activate the concepts to which they are semantically related and vice versa. The activation of a node causes the activation of nodes to which it is connected by excitory links and the deactivation of those to which it is connected by inhibitory links (i.e., competing senses of the same word). Véronis and Ide (1990) built a neural network from the dictionary definitions of the *Collins English Dictionary*. They connect words to their senses and each sense to words occurring in their textual definition. Recently, Tsatsaronis *et al.* (2007) successfully extended this approach to include all related senses linked by semantic relations in WordNet. In several studies, neural networks have been shown to perform well compared to other supervised methods (Leacock *et al.*, 1993; Towell and Voorhees, 1998; Mooney, 1996). However, these experiments are often performed on a small number of words. Major drawbacks of neural networks are the difficulty in interpreting the results, the need for large sets of training data, and the tuning of parameters such as thresholds, decay, etc.

Exemplar-Based / Instance-Based Learning: *Exemplar-based* (or *instance-based*, or *memory-based*) learning is a supervised algorithm in which the classification model is built from *examples*. The model retains examples in memory as points in the feature space and, as new examples are subjected to classification, they are progressively added to the model. An example of such an approach is the k-Nearest Neighbor (kNN) algorithm, one of the highest-performing methods in WSD (Ng, 1997; Daelemans *et al.*, 1999). Currently, exemplar-based learning approaches achieve the best performance in WSD (Escudero *et al.*, 2000a; Fujii *et al.*, 1998; Ng and Lee, 1996; Hoste *et al.*, 2002; Decadt *et al.*, 2004). According to Daelemans *et al.* (1999), these approaches tend to be superior because they do not neglect exceptions and they accumulate further aid for disambiguation as new examples are available.

Support Vector Machines (SVM): introduced by Boser *et al.* (1992), SVM is based on the idea of learning a linear hyperplane from the training set that separates positive examples from negative examples. The hyperplane is located in that point of the hyperspace which maximizes the distance to the closest positive and negative examples (called *support vectors*). SVM has been applied to WSD (Escudero *et al.*, 2000b; Murata *et al.*, 2001; Keok and Ng, 2002; Lee *et al.*, 2004; Buscaldi *et al.*, 2006; Novischi *et al.*, 2007) and has been shown to perform best compared to several supervised approaches.

Ensemble Methods are combination strategies that put together learning algorithms of different nature, that is, with significantly different characteristics. Single classifiers can be combined with different strategies, such as *majority voting*, *probability mixture*, *rank-based combination*, and *adaptive boosting* (*AdaBoost*)(Freund and Schapire, 1999). Ensemble methods are becoming more and more popular as they allow one to overcome the weaknesses of single supervised approaches. When employed on a standard test set, such as that of the Senseval-3 all-words WSD task, ensemble methods overcome state-of-the-art performance among unsupervised systems (up to +4% accuracy) (Mihalcea and Edmonds, 2004).

Minimally and Semi-Supervised Disambiguation

There is not always a clear line separating supervised and unsupervised disambiguation. There also exist minimally or semi-supervised methods which learn sense classifiers from annotated data with minimal or partial human supervision. Such approaches can be based, for example, on automatic bootstrapping of a corpus from a small number of manually tagged examples or on the use of monosemous relatives:

- **Bootstrapping**: a sense classifier is built with little training data, thus overcoming the main problems of supervision, namely the lack of annotated data and the data sparsity problem. Bootstrapping usually starts from few annotated data A, a large corpus of unannotated data U, and a set of one or more basic classifiers. As a result of iterative applications of a bootstrapping algorithm, the annotated corpus A grows increasingly and the untagged dataset U shrinks until some

threshold is reached for the remaining examples in U. The small set of initial examples in A can be generated from hand-labeling (Hearst, 1991) or from the automatic selection with the aid of accurate heuristics (Yarowsky, 1995). The objective of bootstrapping is labeling data which is costly or hard to obtain with no human intervention, by excluding the initial selection of manually annotated examples. There are two main approaches to bootstrapping in WSD, namely *co-training* and *self-training*. The difference between the two is that co-training alternates two classifiers, whereas self-training uses only one classifier which retrains on its own input at each iteration.

Yarowsky (Yarowsky, 1995) and Mihalcea (Mihalcea, 2004) have used the "self-learning" and "co-learning" approaches for WSD, respectively. These methods were based on classifier(s) trained on a small amount of manually tagged data. The same classifiers were then used to tag new data and the most confident predictions were added to the labeled dataset. Yarowsky achieved an accuracy of 96.5% on a test set of 12 ambiguous words with an average of 4000 instances per word. Mihalcea used the same approach on the Senseval-2 generic English corpus and resulted in an improvement of 9.8% over the baseline score using a Bayesian classifier.

A major drawback of co- and self-training is the lack of a method for selecting optimal values for parameters like the pool size p, the number of iterations and the number of most confident examples (Ng and Cardie, 2003). One of the main points of boostrapping is the selection of unlabeled data to be added to the labeled data set.

- **Monosemous Relatives** are possibly synonymous words with a unique meaning. Viewing the Web as a corpus (Kilgarriff and Grefenstette, 2003) is an interesting idea exploited to build annotated data sets, with the aim to relieve the problem of data sparseness in training sets. Such a large corpus can be annotated with the aid of monosemous relatives by way of a bootstrapping algorithm similar to Yarowsky's (Yarowsky, 1995), starting from a few number of seeds. As a result, the automatically annotated data can be used to train WSD classifiers.

In the biomedical domain, a lot of approaches have used supervised machine learning (ML) for WSD. Hatzivassiloglou *et al.* (2001) developed an automated system for assigning protein, gene and mRNA labels to free text. They used three ML techniques, namely naive Bayesian learning, decision trees and inductive rule training and investigated the contribution of different features of textual information (like stopword removal, stemming, positional information of surrounding words) with final accuracy rates up to 85%. Ginter *et al.* (2004) worked on the disambiguation between gene and protein symbols, by introducing a new family of classifiers based on ordering and weighting of the feature vectors obtained from word counts and word co-occurrence in text. This method achieved 86.5% accuracy. Liu et al. (Liu *et al.*, 2002, 2004) showed that there is a need for a larger window size for disambiguation of words in the biomedical domain. Liu *et al.* (2002) used UMLS (Bodenreider, 2004) as ontology. They identified UMLS concepts in abstracts and analyzed the co-occurrence of these terms with the term to be disambiguated. The correct sense was inferred from the majority sense associated with the co-occurring UMLS terms. Co-occurrence was defined using a Bayes approach. The authors achieved a precision of 93% and a recall of 47%. Gaudan *et al.* (2005) used SVMs on their algorithm to resolve abbreviations in MEDLINE and obtained a precision of 98.9% and a recall of 98.2%. Excluding rare senses (appearing in less than 40 documents) from the test set and keeping in the training set only the ambiguous short-forms that also had long-forms in the documents made the disambiguation task easier. Pahikkala *et al.* (2005) followed a similar approach with Schijvenaars *et al.* (2005). But instead of using the full abstract, they defined the context of a gene symbol as a number of words before and after. The size of the context could be varied and optimized. The context was represented as a vector and a support vector machine was trained. They achieved 85% accuracy. Support vector machines are widely used in word sense disambiguation. Their performance depends on a number of parameters such as the sample size, sense distribution and degree of difficulty (Xu *et al.*, 2006). Small datasets and clear or fuzzy borderline between senses impact on the

classification task.

Unsupervised Machine Learning WSD Approaches

Coming to the approaches using unsupervised ML, they have the potential to overcome the *knowledge-acquisition bottleneck* (Gale *et al.*, 1992a), that is, the lack of large-scale resources manually annotated with word senses. These approaches are based on the idea that the same sense of a word will have similar neighbouring words. They use *context-clustering*, *word clustering*, or *co-occurrence graphs*. In context-clustering, each occurrence of a target word in a corpus is represented as a *context-vector*. The vectors are then clustered into groups, each identifying a sense of the target word. In word clustering, words that are semantically similar - and can therefore convey a specific meaning - are clustered together. In a co-occurrence graph $G = (V, E)$, the vertices V correspond to words in a text and edges E connect pairs of words which co-occur in a syntactic relation, in the same paragraph, or in a larger context.

Schütze (Schütze and Pedersen, 1995; Schütze, 1998) adapted LSA/LSI (Latent Semantic Analysis/Indexing) to represent entire contexts rather than single word types using second-order co-occurrences of lexical features. Pedersen and Bruce's (Pedersen and Bruce, 1997, 1998) work with average link clustering relied on a small number of first-order features to create matrices that show the pairwise similarity between contexts. These features were localized around the target word and included word co-occurrences and PoS tags. Purandare and Pedersen (Purandare and Pedersen, 2004) have tested a variety of similar algorithms obtaining an average F-measure of 44%. Yarowsky's (Yarowsky, 1995) and Mihalcea's (Mihalcea, 2004) approaches with "self-learning" and "co-learning" could also be accounted as unsupervised, since the classifier(s) were trained on a small amount of manually tagged data (See Minimally and Semi-Supervised Disambiguation above). Dorow's approach (Dorow and Widdows, 2003) was based on a graph model representing words and relationships (co-occurrences) between them. Sense clusters were iteratively computed by clustering the local graph of similar words around an ambiguous word. The ambiguous words were identified by looking at the nodes connecting otherwise unrelated clusters. These clusters represented the different senses of the word and then the labels were assigned according to WordNet (Fellbaum, 1998), a dictionary of terms and their definitions.

Knowledge-Based WSD Approaches

The objective of knowledge-based WSD is to exploit knowledge resources (such as dictionaries, thesauri, ontologies, collocations, etc. described earlier) to infer the senses of words in context. The main knowledge-based techniques are the *overlap of sense definitions*, *selectional preferences*, and *structural approaches*, like semantic similarity measures and graph-based methods. Most approaches exploit information from WordNet or other knowledge resources. A review of knowledge-based approaches can be found in Manning and Schütze (1999) and Mihalcea (2006).

A simple and intuitive knowledge-based approach relies on the calculation of the word overlap between the sense definitions of two or more target words. This approach is named *gloss overlap* or the *Lesk algorithm* (Lesk, 1986). Given a two-word context (w_1, w_2), the senses of the target words whose definitions have the highest overlap (i.e., words in common) are assumed to be the correct ones. However, the accuracy of such methods remains low, with 18.3% for the original Lesk algorithm and 34.6% for the extended Lesk (Banerjee and Pedersen, 2003).

A historical type of knowledge-based algorithm is one which exploits *selectional preferences* to restrict the number of meanings of a target word occurring in context. Selectional preferences or restrictions are constraints on the semantic type that a word sense imposes on the words with which it combines in sentences (usually through grammatical relationships). For instance, the verb "eat" expects an animate entity as subject and an edible entity as its direct object. We can distinguish between selectional restrictions and preferences in that the former rule out senses that violate the constraint, whereas the latter (more typical of recent empirical work) tend to select those senses which better satisfy the requirements.

In general, approaches to WSD based on selectional restrictions have not been found to perform as well as Lesk-based methods (Navigli, 2009).

Since the availability of computational lexicons like WordNet, a number of *structural approaches* have been developed to analyze and exploit the structure of the concept network in such lexicons. Two main structural approaches are similarity-based and graph-based methods. Similarity-based methods rely on different similarity measures that have been developmed since the early 1990s. For more details on semantic similarity measures and some of their applications, see Section 2.2.2. Graph-based approaches exploit the graph structures to determine the most appropriate senses for words in context. Most of them are related to the notion of *lexical chain* (Halliday and Hasan, 1976; Morris and Hirst, 1991). A lexical chain is a sequence of semantically related words $w_1, ..., w_n$ in a text, such that w_i is related to w_{i+1} by a lexicosemantic relation (e.g., is-a, has-part, etc.). Mihalcea *et al.* (2004) presented an approach based on the use of the PageRank algorithm (Brin and Page, 1998) to study the structure of the lexicon network and identify those nodes (senses) which are more relevant in context. Navigli and Velardi (2005) recently proposed the Structural Semantic Interconnections (SSI) algorithm, a development of lexical chains based on the encoding of a context-free grammar of valid semantic interconnection patterns. A key feature of the algorithm is that it outputs justifications for sense choices in terms of semantic graphs which can be used as a support for the validation of manual and automatic sense annotations. SSI outperformed state-of-the-art unsupervised systems in the Senseval-3 all-words and the Semeval-2007 coarse-grained all-words competition.

As far as the biomedical domain is concerned, there have been developed some knowledge-based approaches, especially in the problem of gene/protein symbols' abbreviations. Wren et al. (Wren *et al.*, 2005) presented a collection of four databases maintaining a vast list of abbreviations together with their meaning. Schijvenaars et al. (Schijvenaars *et al.*, 2005) and Pahikkala et al. (Pahikkala *et al.*, 2005) developed two approaches to resolve gene/protein symbols. Schijvenaars et al. (Schijvenaars *et al.*, 2005) achieved 92.5% accuracy on human gene symbols. The authors compared a gene's definition compiled from a database to abstract where the gene symbol occurs. Both definition and abstract were represented as concept finger prints, i.e., vectors of biomedical terms. Both vectors were compared by a similarity measure based on cosine. Humphrey et al. (Humphrey *et al.*, 2006) at the NLM used lately the Journal Descriptor Indexing (JDI) methodology to handle the ambiguity problem when trying to map free text to terms from the UMLS metathesaurus. JDI combined a statistical, corpus-based method with utilization of pre-existing medical domain knowledge sources. For the 45 ambiguities studied, the overall average precision of the highest-scoring JDI method was 78.7% compared to 25% for their baseline method based on the frequency counts of MeSH terms in a document subset.

2.1.2 WSD Approaches in the Biomedical Domain

As already mentioned in Chapter 1, in the biomedical domain the focus has been on supervised (Hatzivassiloglou *et al.*, 2001; Liu *et al.*, 2004; Gaudan *et al.*, 2005; Pahikkala *et al.*, 2005) and knowledge-based approaches (Schijvenaars *et al.*, 2005; Humphrey *et al.*, 2006; Hakenberg *et al.*, 2008; Farkas, 2008) to disambiguation. These approaches use cosine similarity (Schijvenaars *et al.*, 2005), SVM (Gaudan *et al.*, 2005; Pahikkala *et al.*, 2005), Bayes, decision trees, induced rules (Hatzivassiloglou *et al.*, 2001), and background knowledge sources such as the Unified Medical Language System (UMLS)(Bodenreider, 2004), Medical Subject Headings (MeSH)(Nelson *et al.*, 2001), and the Gene Ontology (GO)(Ashburner *et al.*, 2000). Two approaches use metadata, such as authors (Farkas, 2008) and Journal Descriptor Indexing (Humphrey *et al.*, 2006). Most of the *unsupervised* approaches so far were evaluated outside the biomedical domain (Schütze and Pedersen, 1995; Schütze, 1998; Pedersen and Bruce, 1998; Purandare and Pedersen, 2004; Yarowsky, 1995; Dorow and Widdows, 2003; Mihalcea, 2004), with the exception of (Widdows *et al.*, 2003), who used relations between terms given by the UMLS for unsupervised WSD of medical documents and achieved 74% precision and 49% recall. Another approach that uses the UMLS

as background knowledge for WSD is that of (Leroy and Rindflesch, 2005), who compared the results from a naive Bayes classifier and other algorithms (decision tree, neural network) to conclude that different senses in the UMLS could contribute to inaccuracies in the gold standard used for training, leading to varied performance of the WSD techniques. Another approach by (Dorow and Widdows, 2003) is based on a graph model representing words and relationships (co-occurrences) between them and uses WordNet (Fellbaum, 1998) for assigning labels.

Interestingly, most of the above approaches consider the background knowledge sources as terminologies, without taking into account the taxonomic structure or the terms' semantic similarity (Rada *et al.*, 1989; Sussna, 1993; Resnik, 1995; Lin, 1998; Lord *et al.*, 2003; Azuaje *et al.*, 2005; Schlicker *et al.*, 2006; del Pozo *et al.*, 2008) (see Section 2.2.2 below). Our hypothesis is that ontologies can improve disambiguation. In the next sections we review taxonomies, thesauri and ontologies in the life sciences, some of which we later use to perform disambiguation.

2.2 Ontologies, Text mining and WSD

2.2.1 Ontologies in the Life Sciences

Taxonomies, Ontologies, Thesauri are all background knowledge resources which are related but differ in their degree of expressiveness and support for reasoning.

A *taxonomy* is a form of a classification scheme, arranged in a hierarchical structure, organized by supertype-subtype relationships, also called parent-child relationships. In a taxonomy, children (subtypes) inherit all the properties and constraints of their parents (supertypes) and can have additional ones. One of the best known forms of taxonomies is the "Linnaean taxonomy", a biological classification of organisms.

A *thesaurus* is a type of controlled vocabulary used mainly for indexing or tagging purposes. A thesaurus groups together terms that are semantically close to each other. The relationships between the terms in a thesaurus can be hierarchical ('broader than', 'narrower than'), equivalent (to connect synonyms and near-synonyms, e.g., 'used for') and associative, used to connect related terms whose relationship is neither hierarchical nor equivalent (e.g., 'related to').

A common definition for an *ontology* is "a formal explicit specification of a shared conceptualization" (Gruber, 1993). According to Tim Berners-Lee, "an ontology is a document or file that formally defines the relations among terms. The most typical kind of ontology for the Web has a taxonomy and a set of inference rules" (Berners-Lee *et al.*, 2001). An ontology is a formal representation of a set of concepts within a specific domain and the relationships between them. Within an ontology, the types of relations between the concepts can be more than simple supertype-subtype, therefore ontologies are broader and more flexible than taxonomies and thesauri. Ontologies provide dynamic, controlled vocabularies of concepts to help manage the interoperability between data sources. A typical ontology is a hierarchical structure of concepts (classes), definitions for these concepts, and associations between concepts. Additional logical axioms serve as further constraints among these entities. In a state-of-the art setting, an agent queries the ontology and a knowledge base that is based on this ontology. By exploiting the structure of an ontology, specific and reliable retrieval becomes possible.

At present the field of biology also faces the problem of the presence of a large amount of data without any associated semantics. Therefore, biologists currently waste a lot of time and effort in searching for all of the available information about each small area of research. This is hampered further by the wide variations in terminology that may be in common usage at any given time, and that inhibit effective searching by computers as well as people.

In recent years, to facilitate biomedical research, various ontologies and knowledge bases have been developed. For example the Gene Ontology (GO) project is a collaborative effort to address the need for consistent descriptions of gene products in different databases (Ashburner *et al.*, 2000). Another widely

used system has been developed by the United States National Library of Medicine called the Unified Medical Language System (UMLS) which is a consolidated repository of medical terms and their relationships, spread across multiple languages and disciplines (chemistry, biology, etc) (Bodenreider, 2004). Medical Subject Headings (MeSH) is a controlled vocabulary maintained by the U.S. National Library of Medicine[12], mainly used for annotating and indexing articles from PubMed (Nelson *et al.*, 2001). Moreover, several specialised databases for various aspects of biology have been developed, providing rich vocabularies including several synonyms. For example, the UniProt/Swiss-Prot Knowledge base[13] is an annotated protein sequence database.

Gene Ontology

A renowned ontology in the life sciences, especially for biology and bioinformatics, is the Gene Ontology (GO) (Ashburner *et al.*, 2000). It provides a controlled vocabulary to annotate gene products according to the biological processes in which they participate, the molecular functions they perform, and the cellular location in which they act. With each resource describing its gene products in a common form, this sharing, together with the structure provided by the relationships between terms in the GO, makes querying of within and between resources possible. A section of the GO graph is given in Figure 2.7

Although the Gene Ontology was created for the express purpose of providing a common terminology for functional annotation of genes and gene products in biological databases towards the goal of database interoperability, it has since been widely used for a variety of purposes, including analyses of experimental data, predictions of experimental results, and document retrieval.

GO is the most prominent ontology of the **Open Biomedical Ontologies (OBO)**[14], a collection of biological ontologies that are open in that they can be used by all without constraint so long as the sources are acknowledged and the ontologies are not edited and re-distributed under the same names. In addition to the taxonomies of GO, the OBO ontologies deal with anatomies of humans and of various model organisms, biochemical substances, and sequence types, among others. Over the past years, GO has developed into the main ontology in molecular biology. Today, over 19,000 terms organised in three sub-ontologies (biological process, molecular function, cellular location) comprise the Gene Ontology. The terms are linked by three relations, 'is–a', 'part–of', 'is–synonym'.

Medical Subject Headings (MeSH)

The Medical Subject Headings (MeSH) is a controlled vocabulary maintained by the U.S. National Library of Medicine[15] (Nelson *et al.*, 2001). It is mainly used for annotating and indexing articles from PubMed. The MeSH terminology provides a consistent way to retrieve information that may use different terminology in different articles for the same concepts. In the 2009 MeSH there are 25,186 descriptors, with an additional over 160,000 supplementary concepts, called entry terms. These entry terms assist in finding the most appropriate MeSH Heading, for example, "Vitamin C" is an entry term to "Ascorbic Acid". In addition to these headings, there are more than 180,000 headings called Supplementary Concept Records within a separate thesaurus. MeSH is organized in a tree, with concepts such as anatomy and diseases, but also geographic locations, at the top level. The MeSH vocabulary is used for indexing journal articles from Index Medicus and Medline and also for cataloguing books and audiovisuals. PubMed contains links to full-text articles at participating publishers' websites as well as links to other third party sites. It also provides access and links to the integrated molecular biology databases maintained by the National Center for Biotechnology Information. Table 2.2 shows the main differences between GO and MeSH.

[12] See http://www.nlm.nih.gov/mesh/meshhome.html
[13] See http://www.uniprot.org/
[14] See OBO foundry http://www.obofoundry.org/
[15] See http://www.nlm.nih.gov/mesh/meshhome.html

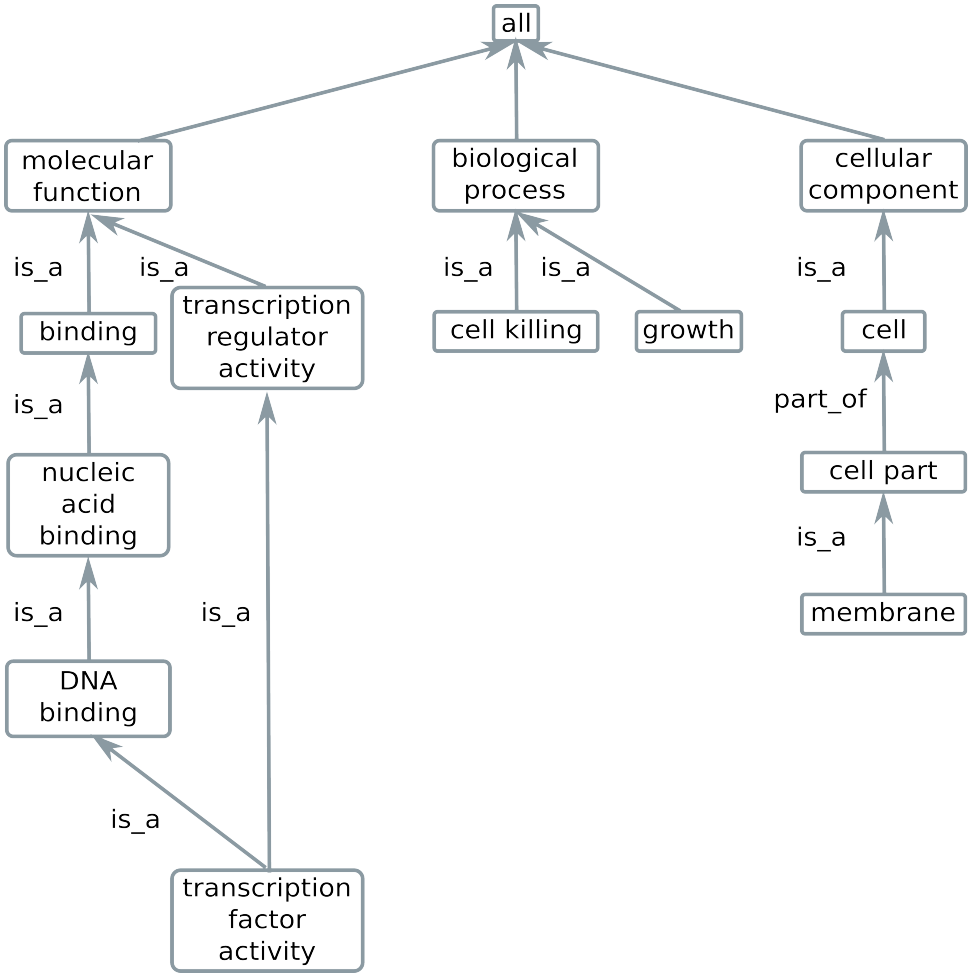

Fig. 2.7: Section of the GO graph showing the three aspects (molecular function, biological process, and cellular component) and some of their descendant terms. The fact that GO is a directed acyclic graph (DAG) rather than a tree is illustrated by the term 'transcription factor activity' which has two parents. An example of a *part_of* relationship is also shown between the terms 'cell part' and 'cell'.

	Gene Ontology (GO)	**Medical Subject Headings (MeSH)**
Primary purpose	gene product annotation (biological process, molecular function, cellular location)	annotation & indexing of biomedical articles (Index Medicus, MEDLINE)
Number of concepts	>19,000 terms	25,186 descriptors, >160,000 entry terms
Type of relations	'is–a', 'part–of', 'is–synonym'	*A narrower than B* (so that users interested in *B*s are given the option to look at *A*s), associative relationships (see http://www.nlm.nih.gov/mesh/intro_entry.html)

Tab. 2.2: Gene Ontology (GO) vs. Medical Subject Headings (MeSH).

Unified Medical Language System (UMLS)

The Unified Medical Language System[16] (UMLS) is an attempt to make a collection of clinical/medical terminologies interoperable (Bodenreider, 2004). The UMLS Metathesaurus is a large, multi-purpose, and multilingual vocabulary database that contains information about biomedical and health related concepts, their various names and the relationships among them. It is built from the electronic versions of many different thesauri, classifications, code sets, and lists of controlled terms used in patient care, health services billing, public health statistics, indexing and cataloging biomedical literature, and /or basic, clinical, and health services research. The 2009 release of the UMLS Metathesaurus[17] comprises of 150 biomedical vocabularies, including the Gene Ontology, the Medical Subject Headings, the Foundational Model of Anatomy (FMA) (Rosse and Mejino, 2003), the Systematized Nomenclature of Medicine (SNOMED) (Spackman, 2004) and others. The Metathesaurus does not represent a comprehensive ontology of biomedicine or a single consistent view of the world (except at the high level of the semantic types assigned to all its concepts). It preserves the many views of the world present in its source vocabularies because these different views may be useful for different tasks.

Ontologies for Anatomy

While most biological databases contain information at the molecular level, there is a growing need to link this information to concepts about the *global* structure of organisms, that is to their anatomy. This development is due to two main reasons.

A central question in genetics is which genes influence the development of which parts of an organism and which genetic mutations cause which deviations from the standard phenotype. Researchers tackle this question by exploring which genes are expressed at which stage of development in which tissues of an organism. To make such findings generally accessible, a standardized vocabulary about developmental stages and tissues is needed for annotations. A second reason is that biological image data are increasingly being published on the Web. To describe in a uniform way what tissue an image shows one has to resort to some anatomical vocabulary.

Much data is collected on a variety of organisms, and very often represented in structured, thus queryable, databases:

- *Mouse Genome Informatics* (MGI)[18] gives integrated access to various types of genetic and genomic data on the mouse (Ringwald *et al.*, 2001).
- *Wormbase*[19] has information on the worm C. elegans and other nematodes (Stein *et al.*, 2001).
- *Wormatlas*[20] is another resource for C. elegans and provides an anatomy handbook (Altun and Hall, 2006).
- *FlyBase*[21] collects genomic information on the fruit fly Drosophila (Consortium, 1998).
- *Saccharomyces Genome Database* (SGD)[22] collects genomic information on the baker's yeast, S. cerevisiae.
- *Zebrafish Information Network* (ZFIN)[23] makes gene expression, mutant, and other genomic data on the zebrafish available (Sprague *et al.*, 2006).

[16] See http://www.nlm.nih.gov/research/umls/
[17] See http://www.nlm.nih.gov/research/umls/knowledge_sources/metathesaurus/release/source_vocabularies.html
[18] http://www.informatics.jax.org/
[19] http://www.wormbase.org/
[20] http://www.wormatlas.org/
[21] http://flybase.bio.indiana.edu/
[22] http://www.yeastgenome.org/
[23] http://zfin.org/

Anatomy ontologies can be sizable. The mouse anatomy, for instance, comprises more than 8,000 terms. Anatomies can be integral parts of larger ontologies or controlled vocabularies, like in the Medical Subject Headings (MeSH) system described earlier. MeSH contains mostly terms for human anatomy, but also some that relate to various mammal species. MeSH terms are used to annotate entries in large bibliographical databases.

The Edinburgh Mouse Atlas Project (EMAP) provides a resource that combines an anatomy ontology with a three-dimensional spatial model of the mouse embryo to give access to gene expression data (Baldock *et al.*, 2003). Anatomical terms are linked to regions in the spatial model and vice versa. The Mouse Atlas is based on the same anatomy as Jackson Lab's MGI, but has been enriched it to represent groupings between tissues such as the "skin" group, which comprises tissues in many different locations (Bard *et al.*, 1998). We will elaborate on ontologies for anatomy in Chapter 5, while describing MousePubMed, a system for mining scientific literature on mouse anatomy.

2.2.2 Semantic Similarity of Terms: Measures and Applications

This section provides an overview on semantic similarity measures and some of their applications in biomedical context. So far, most of the aforementioned Word Sense Disambiguation approaches consider the background knowledge sources as terminologies, without taking into account the taxonomic structure or the terms' semantic similarity. This gap is filled by the systematic comparison of the three approaches that use ontologies with inference and semantic similarity and the use of metadata to solve the problem of WSD for ontological terms (see Chapter 3, Section 3.3). In this context, a **semantic similarity measure** is a function that, given two ontology terms, returns a numerical value reflecting the closeness in meaning between them.

Semantic Similarity Measures

An overview of some semantic similarity measures proposed to assess the conceptual distance between concepts or sets of concepts and some of their applications are given in Table 2.3.

Rada *et al.* (1989) were among the first ones to talk about similarity between concepts on Semantic Nets. They proposed a metric, called Distance, in order to assess the conceptual distance between sets of concepts when used on a semantic net of hierarchical relations. Distance between two concepts in a hierarchy is defined as the minimum number of edges separating the concepts. They also defined the Distance on sets of nodes, in order to check the similarity between sets of concepts. They tested the appropriatness of the metric for measurement of the conceptual distance between concepts in MeSH (Nelson *et al.*, 2001) and compared it to human assessment. They conclude that Distance is a valuable tool for simulating human assessments of conceptual distance and evaluating some cognitive aspects of semantic nets. Their long-term goal is to solve the problem of document ranking in response to a query.

Sussna (1993) used the WordNet semantic network (Fellbaum, 1998) and applied disambiguation on a Times magazine corpus (of 5 documents). Sussna introduced the idea of *mutual constraint* among terms and its special case, the *frozen past* approach, in order to achieve total distance minimization (or "energy minimization"). He used a *moving window* of terms in focus while moving from the beginning of a document towards its end. In the *frozen past* approach actually all terms except the one being disambiguated have had their senses determined and "frozen". Sussna concludes that using the moving frozen past window gives ascending performance to a point and then plateaus. The method trades off space for time, with the use of large data structures kept in memory, a minimum runtime processing effort and without any syntactic analysis.

Resnik (1995) introduced and quantified a new measure for semantic similarity, the *information content* of a concept. He converted the measure from pure distance (number of intervening is_a links) to similarity. Resnik defined the similarity between two concepts as the extent to which they share information in common. Considering this in a hierarchical concept/class space, this common information

"carrier" could be identified as a specific concept node that subsumed both of the two in the hierarchy (a parent super-class of both). The similarity value was defined as the information content value of this specific super-ordinate class. The value of the information content of a class was then obtained by estimating the probability of occurrence of this class in a large text corpus. The problem is that, sometimes, the measure produces fake high similarity measures for words on the basis of inappropriate word senses (e.g., due to synonyms). In measuring similarity between words, it is the relationship among word senses that matters.

Richardson and Smeaton (1995) introduced an approach to Information Retrieval (IR) based on computing a semantic distance measurement between concepts of words and using this word distance to compute a similarity between a query and a document. They applied Resnik's (Resnik, 1995) *information-based* similarity estimator and Rada's *conceptual distance* estimator to WordNet synsets and found that the measures were less accurate than expected. Richardson and Smeaton found that irregular densities of links between concepts result in unexpected conceptual distance outcomes.

Jiang and Conrath (1997) proposed a combined approach that inherits the edge-based approach of the edge counting scheme (Rada's distance (Rada *et al.*, 1989)), enhanced by the node-based approach of the information content calculation (Resnik's information content (Resnik, 1995)). They first considered the *link strength* factor which is the difference of the information content values between a child concept and its parent concept. Considering other factors, such as *local density* (the greater the density, the closer the distance between the nodes), *node depth* (distance shrinks as one descends the hierarchy) and *link type* (relation type, is_a, part_of), Jiang and Conrath first defined the overall edge weight for a child node and its parent and then the overall distance between two nodes as the summation of edge weights along the shortest path linking the two nodes. Jiang and Conrath tested their approach on a common dataset of word pair similarity ratings, outperformed other computational models and gave the highest correlation value with a benchmark based on human similarity judgements.

Lin (1998) provided a universal definition of similarity in terms of information theory: "the similarity between A and B is measured by the ratio between the amount of information needed to state the commonality of A and B and the information needed to fully describe what A and B are". Lin demonstrated the universality of this definition by its application in different domains, such as similarity between ordinal values, feature vectors (string similarity), word similarity and semantic similarity in a taxonomy.

Approaches to measuring semantic similarity (or *semantic relatedness*) can be categorized into *dictionary-based*[24], *corpus-based* and *hybrid* (Budanitsky and Hirst, 2006; Tsatsaronis *et al.*, 2010). Resnik's measure (based on the Information Content, Resnik (1995)) can be considered as a hybrid measure, since it combines both the hierarchy of the used thesaurus and statistical information for concepts measured in large corpora. The same applies for the measures of Jiang and Conrath (1997) and Lin (1998).

Budanitsky and Hirst (2006) performed an evaluation of five semantic similarity measures (Jiang and Conrath, 1997; Hirst and St-Onge, 1998; Leacock and Chodorow, 1998; Lin, 1998; Resnik, 1995), all of which use WordNet as their central resource, by comparing their performance in detecting and correcting real-word spelling errors. The information-content-based measure proposed by Jiang and Conrath (Jiang and Conrath, 1997) was found to perform best.

Applications in Biomedical Ontologies

Lord *et al.* (2003) implemented GOGraph, a tool for calculating the semantic similarity of protein pairs based on Resnik's information content measure. They investigated the application of semantic similarity measures to ontological annotations of the SWISS-PROT database, as well as how the ontological structure affects the similarity.

[24] Also found in the bibliography as knowledge-based, thesaurus-based, or lexicon-based.

Metric	Description
Rada *et al.* (1989)	*min* # of edges separating the concepts
Sussna (1993)	*frozen past*: all terms except the ambiguous one have their senses determined & frozen
Resnik (1995)	*information content* (common information between two concepts)
Lin (1998)	universal definition: $sim_{(A,B)} = \frac{info\ needed\ to\ state\ the\ commonality\ of\ A\ and\ B}{info\ needed\ to\ fully\ describe\ what\ A\ and\ B\ are}$ (application as *sim* between ordinal values, feature vectors, word *sim* & semantic *sim* in taxonomy)
Application	**Description**
Richardson and Smeaton (1995)	Resnik + Rada, measure distance between concepts of words to compute *sim* between a query and a document (Information Retrieval)
Jiang and Conrath (1997)	Rada + Resnik, compared to human similarity judgements
Lord *et al.* (2003)	Resnik, semantic *sim* of protein pairs
Azuaje *et al.* (2005)	gene similarity (GO terms assigned)
Schlicker *et al.* (2006)	Lin + Resnik, comparison of sets of GO terms & gene functional *sim* assessment
Camous *et al.* (2007)	Resnik, sem *sim* in MeSH; extend MeSH representation of Medline docs
del Pozo *et al.* (2008)	functional distance between GO terms (term cooc in Interpro)

Tab. 2.3: Semantic similarity measures and some applications.

Azuaje *et al.* (2005) used a semantic similarity measure to assess gene similarity with a view to providing a solid basis for the implementation of classification tools and the automated validation of functional associations. Azuaje et al. assessed the similarity between genes based on their GO terms. They used the *distance* measure and considered only the best semantic match amongst genes of group B for each gene in group A. The method gave an asymmetrical measure expressing the semantic contribution of A genes in relation to B.

Schlicker *et al.* (2006) introduced two semantic similarity measures for comparing sets of GO terms and for assessing the functional similarity of gene products. The first measure (sim_rel) was based on Lin's (Lin, 1998) and Resnik's (Resnik, 1995) measures and took into account how close two GO terms are to their lowest common ancestor (LCA) as well as the LCA's relevance (i.e. how general/specific it is). Based on the sim_rel score, the second measure, called funSim, compared the annotation of two gene products. The funSim score could compare two sets of GO terms from different ontologies and allowed for partial matches (was independent from the sequence similarity). Therefore, it was suitable for comparison of multi-functional gene products.

Camous *et al.* (2007) applied Resnik's information content measure to evaluate semantic proximity between concepts within the MeSH hierarchy. They proposed a method for extension of ontology-based representations of biomedical documents and used the Medical Subject Headings for this representation. The initial MeSH-only representations were extended with MeSH concepts that were semantically close within the MeSH hierarchy. The extension method was evaluated within a document triage task organized by the Genomics track of the 2005 Text REtrieval Conference (TREC) and lead to an improvement of 18.3% over a non-extended baseline in terms of normalized utility, the metric defined for the task.

A recent review by Pesquita *et al.* (2009) describes the semantic similarity measures applied to biomedical ontologies and proposes a classification according to the strategies they employ: *node-based* vs. *edge-based* and *pairwise* vs. *groupwise*. The authors also survey the existing implementations of semantic similarity measures and describe examples of applications to biomedical research.

2.2.3 Ontology Engineering and Text Mining

The engineering of ontologies is still a new research field. There does not yet exist a well-defined theory and technology for ontology construction. This means that many of the ontology design steps remain manual and a kind of "art" and intuition (Soldatova and King, 2005; Sowa, 2000; Castro *et al.*, 2006). There exists a variety of different ontologies, constructed for different purposes and projects.

As far as the biomedical ontologies are concerned, during the last years there have been major efforts in the biological community for organizing biological concepts in the form of controlled terminologies or

ontologies (Eilbeck *et al.*, 2005; Whetzel *et al.*, 2006; Ashburner *et al.*, 2000; Evsikov *et al.*, 2004). A key difference between terminologies and ontologies is that the former lack the semantic depth of the latter. However, when it comes to design, terminologies can serve as basis for ontologies and vice-versa. An example where a terminology can serve for ontology is that of the Gene Ontology (Ashburner *et al.*, 2000), which provides a controlled vocabulary to describe gene and gene products in any organism. On the other side, the Gene Ontology Next Generation (GONG) project (Wroe *et al.*, 2003) aims at the migration of current bio-ontologies to a richer and more rigorous status, using formal representation languages like OWL. Examples of true ontologies are the GALEN project (Rector *et al.*, 1996) and the Systematized Nomenclature of Medicine (SNOMED) (Spackman, 2004) which are based on Description Logic for concept representation and the Foundational Model of Anatomy (FMA) (Rosse and Mejino, 2003) which is based on frames representing information about anatomical classes, designed so that content can be maintained as a dynamic resource and can be used as terminologies.

There have also been developed systems to provide interoperability among different ontologies, such as the Unified Medical Language System (Bodenreider, 2004) in order to provide a common frame of reference among the different research communities. The **Open Biomedical Ontologies (OBO) Foundry**[25] hosts over 60 open source ontologies associated with phenotypic and biomedical information, such as the Mouse Anatomy (MA) (Evsikov *et al.*, 2004) and the Cell Ontology (CL) (Bard *et al.*, 2005). Schulz *et al.* (2007) have recently proposed a formalism that deals with "role propagation along non-taxonomic hierarchies", and appears suitable for the redesign of compositional hierarchies in bio-ontologies of the OBO Relation Ontology framework. Bodenreider and Stevens (2006), Blake and Bult (2006) and Baker *et al.* (1999) give overviews on biomedical ontologies, the consortia involved, formalisms as well as semantic web technologies and representation tools.

Semantic meta-information provided in the form of ontologies has proven useful in order to search (Doms and Schroeder, 2005) or index large collections of documents (e.g., MeSH for indexing MEDLINE (Nelson *et al.*, 2001)). Meta-information found for text documents is often general (keyword list) or still too complex for an automated evaluation (article abstract). Finding terms of controlled vocabularies in text overcomes this shortage, while relations between terms provide the necessary navigation structures.

Ontological background knowledge can serve to answer questions with knowledge-based search engines, by easing the task of finding relevant documents through the term automatic annotation (Doms and Schroeder, 2005; Mueller *et al.*, 2004; Perez-Iratxeta *et al.*, 2003; Wermter *et al.*, 2009; DeLuca *et al.*, 2009). In the domain of lipoprotein metabolism, for example, a search for "analphalipoproteinemia" will retrieve articles for Tangier's disease, which is actually a synonym. In case of a syndrome, such as the "metabolic syndrome", in a properly designed ontology the articles retrieved will contain symptoms and other characteristics for it (e.g., type II diabetes, hypertension, insulin resistant, low HDL, hypertension, all of them being parts of the metabolic syndrome). Researchers explore literature on different parameters that can affect the lipoprotein metabolism, such as the phenotype, genotype and age of the patients/animals tested, environmental factors and lifestyle, specific lipoprotein and enzyme concentrations and others. Questions like 'what is the activity of cholesterol ester transfer protein in diabetes', 'which cells/tissues is apoE expressed in', 'what is the impact of a fish oil diet on metabolic syndrome individuals', 'which genes/proteins/metabolites are hypertension-specific' can be answered with the use of a well designed ontology on lipoprotein metabolism, containing terminology found in literature with semantically interconnected terms.

The **GoPubMed**[26] search engine (Doms and Schroeder, 2005) allows users to explore PubMed search results with the Gene Ontology (GO) (Ashburner *et al.*, 2000) and Medical Subject Headings (MeSH) (Nelson *et al.*, 2001). GoPubMed retrieves PubMed abstracts for a search query, detects terms from the GO and MeSH in the abstracts, displays the subset of GO and MeSH relevant to the keywords and allows for browsing the ontologies and displaying only articles containing specific GO and MeSH terms

[25] See OBO foundry http://www.obofoundry.org/
[26] See http://gopubmed.org/

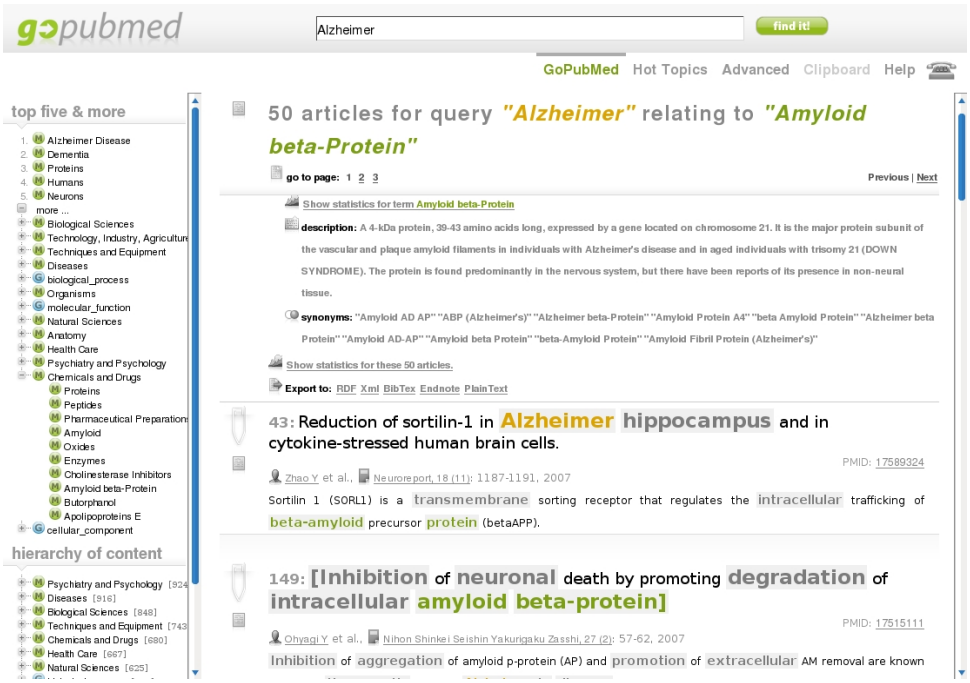

Fig. 2.8: Which proteins are related to Alzheimer's disease? GoPubMed uses its ontological background knowledge to index search results according to the Gene Ontology and MeSH. Abstracts of articles in PubMed are automatically annotated with GO/MeSH terms and protein names and the results of the search are grouped based on these annotations.

(see Figure 2.8 for an example). The search engine is developed in a way that any ontology (e.g., a Lipoprotein Metabolism Ontology) can be easily integrated and used for a domain-specific literature search. One of the benefits of such an ontology-based literature search is the categorization of abstracts according to a specific ontology, allowing users to quickly navigate through the abstracts by category and providing an overview of the literature. It can also automatically show general ontology terms related to the original query, which often do not even appear directly in the abstract.

In Chapter 4 we introduce design principles for ontologies used for text mining, based on our personal experience with the manual development of a Lipoprotein Metabolism Ontology. This LMO ontology was later used together with the GoPubMed infrastructure to assist researchers from Unilever[27] into lipoprotein-metabolism-specific literature search.

[27] http://www.unilever.com/

2.3 The Semantic Web and Semantic Search

Semantic Web The World Wide Web is a collection of documents/pages interconnected by links. Vast amounts of information remain inaccessible, well-hiden into databases that need to be queried by users. The World Wide Web has a huge amount of data and is a reliable source of information for many topics. However, since there is not much semantics associated with that data, the information cannot be processed by autonomous computer agents and is only understandable to humans.

The *Semantic Web* is a vision of the next generation World-Wide Web in which data from multiple sources described with rich semantics are integrated to enable processing by humans as well as software agents. One of the goals of Semantic Web research is to incorporate most of the knowledge of a domain in an ontology that can be shared by many applications. *Ontologies* organise information of a domain into taxonomies of concepts, each with their attributes, and describe relationships between concepts. Mukherjea (2005) gives an overview on semantic web languages and current efforts to represent biomedical knowledge, as well as on techniques that have been developed to effectively retrieve information from the semantic web.

Semantic Annotation In order to make the content of web pages machine-understandable, people use *semantic annotation*. The web pages are being manually annotated with tags that provide a specific sense to text. For example, in a web page about a certain protein, types of information such as the protein name, function, three dimensional structure, location in the cell and source organism can be annotated with XML mark-up. To continue, the marked-up text can be recognized by other programs and exploited. A key point to this annotation is the agreement on the *senses* of the annotations. Such senses can be defined and structured into *ontologies*, where terms can have synonyms and be combined to form new ones.

Semantic Web Browsers The sheer volume of resources available online makes it increasingly harder for users to find specific information and make quality judgements (Roy *et al.*, 2006). This problem is of particular concern to the life sciences, where sharing and making data available on the Web is widely accepted (Schroeder *et al.*, 2006). Commonly, scientists and medical practitioners need easy access to information about chemical compounds, biological systems, diseases, and the interactions between these entities, which requires this data to be effectively integrated (W3C Interest Group, 2008). The emerging Semantic Web (SW) technology (Berners-Lee *et al.*, 2001) aims to provide a solution. While general purpose Semantic Web Browsers (SWBs) such as Tabulator[28] may enhance the search and browsing experiences of everyday users, Semantic Web technology in the life sciences has the potential to address the urgent needs of clinicians to find specific, quality-assured information under severe pressure of time (Gray and de Lusignan, 1999). Through Semantic Web Browsers, using underlying domain ontologies, context-based knowledge integration and semantically enhanced navigation can be achieved.

Semantic Web languages and formalisms Ontologies may vary in their content, structure and implementation. A number of possible formal languages can be used for ontology construction, including general logic programming languages (like Prolog). More common, however, are languages that have evolved specifically to support ontology construction. When comparing ontology languages, what is given up for computability and simplicity is usually language expressiveness. A language needs only be as rich and expressive as is necessary to represent the nuance and intricacy of knowledge that the ontology's purpose and its developers demand. Several ontology languages have been developed during the last years. Some of them are based on XML syntax, such as Ontology Exchange Language (XOL), SHOE (which was previously based on HTML), and Ontology Markup Language (OML) (using a markup scheme to encode knowledge), whereas Resource Description Framework (RDF) and RDF Schema are

[28] See http://www.w3.org/2005/ajar/tab

languages created by World Wide Web Consortium (W3C) working groups. Finally, two additional languages are being built on top of RDF(S) - the union of RDF and RDF Schema - to improve its features: Ontology Inference Layer (OIL) (Fensel *et al.*, 2001) and DAML+OIL[29]. OWL[30] is a language for making ontological statements, developed as a follow-on from RDF and RDFS, as well as earlier ontology language projects including OIL, DAML and DAML+OIL. OWL is intended to be used over the World Wide Web, and all its elements (classes, properties and individuals) are defined as RDF resources, and identified by URIs (Uniform Resource Identifiers).
The Web Ontology Language (OWL) is the most recent development in standard ontology languages, endorsed by the World Wide Web Consortium (W3C) to promote the Semantic Web vision. An OWL ontology may include descriptions of classes, properties and their instances. Given such an ontology, the OWL formal semantics specifies how to derive its logical consequences, i.e. facts not literally present in the ontology, but entailed by the semantics.

[29]See DAML+OIL http://www.daml.org/2001/03/daml+oil-index.html
[30]See OWL working group at http://www.w3.org/2007/OWL/wiki/OWL_Working_Group

Chapter 3

Word Sense Disambiguation

With more and more genomes being sequenced, a lot of effort is devoted to their annotation with terms from controlled vocabularies such as the Gene Ontology (Ashburner *et al.*, 2000). Manual annotation based on relevant literature is tedious, but automation of this process is difficult. One particularly challenging problem is word sense disambiguation. Ontology term labels such as 'development', 'spindle', 'nucleus', 'cell', 'host' can be ambiguous and have multiple senses. 'Development', for example, can refer to developmental biology or to the more general sense. While this is no problem for human annotators, it is a challenge to automated methods, which identify ontology terms in text.

Classical approaches to word sense disambiguation use co-occurrence analyses of words/terms to the ambiguous term to reach a decision about the correct sense. However, most treat ontologies as simple terminologies, without making use of the ontology/hierarchy structure or the semantic similarity between terms. Another useful source of information for disambiguation are metadata. Examples of metadata in the abstract of a scientific article in PubMed can be the title of the journal, the title of the article, the author name and other features not directly included in the text of the abstract.

The current chapter gives details on four approaches to address the problem of ambiguous biomedical terms, making use of term co-occurrences, document clustering, the ontology structure, the semantic similarity between terms and metadata. The co-occurrence method, called *Term Cooc*, is first compared against *document clustering* which groups documents containing similar ontology terms (see Section 3.2, Term Co-occurrences vs. Document Clustering).

Furthermore, the Term Cooc method is extended by several means (using the hierarchy structure in GO and MeSH, and combination with a support vector machine in a co-training scheme) and a systematic comparison of Term Cooc and two more disambiguation methods developed by partners is performed. The first method, called *MetaData*, uses metadata such as the ones mentioned earlier. The second method, called *Closest Sense*, uses the UMLS semantic network as background knowledge for the computation of similarities between the different senses of the ambiguous term, the senses of its neighbours and the type of relations that could occur between them. A systematic comparison of the *Term Cooc*, *MetaData* and *Closest Sense* disambiguation approaches is performed on a manually curated training corpus for seven ambiguous terms from the Gene Ontology and MeSH (see Section 3.3, Term Cooc vs. Closest Sense vs. MetaData).

The current work on word sense disambiguation has been published in bioinformatics journals (Andreopoulos *et al.*, 2008; Alexopoulou *et al.*, 2009), and presented in several bioinformatics and computational biology conferences as poster (Alexopoulou *et al.*, 2007a,b) or short presentations (Alexopoulou *et al.*, 2008a). It has also been part of the EU–funded project Sealife[1] (Schroeder *et al.*, 2006).

[1] http://www.biotec.tu-dresden.de/sealife/

3.1 Motivation and contribution

Most of the aforementioned Word Sense Disambiguation approaches described in Section 2.1 and Table 2.1 consider the background knowledge sources as terminologies, without taking into account the taxonomic structure or the terms' semantic similarity (Rada *et al.*, 1989; Sussna, 1993; Resnik, 1995; Lin, 1998; Lord *et al.*, 2003; Azuaje *et al.*, 2005; Schlicker *et al.*, 2006; del Pozo *et al.*, 2008) (see Introduction, Section 2.2.2). An exception to that is the work of Resnik (1999), who applied semantic similarity to assign confidence values to word senses of nouns within thesaurus-like groupings. Here, the gap is filled by the systematic comparison of the three approaches that use ontologies with inference and semantic similarity and the use of metadata to solve the problem of WSD for ontological terms (see Section 3.3, Term Cooc vs. Closest Sense vs. MetaData). We investigate means of embedding information from the ontology structure and metadata to improve the performance of word sense disambiguation.

We first develop and evaluate two approaches to the WSD problem, namely ***'Term Cooc'*** (term co-occurrences in PubMed abstracts) and ***document clustering***, with the results published in Andreopoulos *et al.* (2008) (see Section 3.2). We propose a methodology for finding whether the annotation of an article in an automatically annotated database is likely to be true or false with respect to the biological meaning and construct a co-occurrence graph of GO terms based on Gene Ontology Annotations (GOA) (Camon *et al.*, 2004).

We further extend the ***'Term Cooc'*** approach in the following ways: first, we additionally disambiguate MeSH terms and use a larger training corpus to get the co-occurrence scores, since there exist $\sim 16,400,000$ documents to which experts have assigned MeSH terms. Second, we make use of the hierarchy structure in both GO and MeSH (given an ambiguous term α, the co-occurrence of α with a term β should not be lower than α's co-occurrence with any of β's descendants, 'Inferred Cooc'). We therefore investigate how two different hierarchies influence the performance[2] of disambiguation. Third, we combine our graph–based decision function with a support vector machine, arranged in a co–training scheme, to learn and improve models without any labelled data. Finally, we test the disambiguation performance in new larger benchmark datasets of varying curation quality that we collected[3]. The 'Term Cooc' approach is similar to Dorow's approach (Dorow and Widdows, 2003), with the difference that we construct the co-occurrence graph based on GOA and MeSH, which are manually annotated datasets. Therefore, our graphs contain only relations (edges) between terms (nodes) that are semantically meaningful in the context of an article. Dorow's graph contains all the nouns that co-occur, but in the case of the biological context we are interested only in a local subgraph of Dorow's graph (i.e. 'development' only in the biomedical sense). Another difference is that we use established knowledge in GO and MeSH to draw the nodes and in the different configurations of our 'Term Cooc' method we use a support vector machine and/or incorporate the term relationships in GO and MeSH.

We introduce two more methods for disambiguation, differing from the 'Term Cooc' approach in terms of automation and background knowledge required. The ***'Closest Sense'*** approach computes similarities between the senses of the ambiguous term, the senses of its neighbours (co-occurring terms) and the type of relations that could occur between them ('subClassOf' relations as well as 'subPropertyOf' relations, see Section 3.3.1, Closest Sense, Semantic distances). 'Closest Sense' (CS) uses the UMLS semantic network as background knowledge, like Widdows *et al.* (2003), who rely on the context of the ambiguous term in order to compute a score for each sense candidate. This score consists of the number of terms in the document which are related, in the UMLS, with the different senses of the ambiguous term. In comparison to the 'Closest Sense' method, this approach is different in two main points : (i) it does not take advantage of the hierarchies of concepts and relations in the UMLS and (ii) it ignores terms which co-occur with the ambiguous term in the same context but do not have a direct link with it in the

[2]We measure performance in terms of Precision, Recall, Specificity and f–measure.

[3]The benchmarks were manually, semi-automatically and automatically collected, see Section 3.3.2, Datasets, and Appendix A

UMLS. The ***'MetaData'*** method uses maximum entropy for modelling the behaviour of occurrence of contextual terms and phrases in text together with a potentially ambiguous term. The features selected are n-tuples of word stems and metadata such as the journal and document title. The method requires a set of labelled documents for each term to be disambiguated.

We evaluate and compare the three strategies for WSD, namely 'Closest Sense', 'Term Cooc', and 'MetaData', starting from the unsupervised/automated to the least automated one. The comparison includes each method's requirements and limitations in terms of training data and automation, the behaviour of the methods during the use of different taxonomies (GO/MeSH/UMLS) as well as comparison against a classical stem co-occurrence approach. We additionally make the benchmark datasets created for the purpose of disambiguation publicly available, since the collection process is time-consuming and labour-intensive. These include 2600 manually curated documents of high/medium curation quality for 7 selected GO and MeSH terms[4].

3.2 Term Co-occurrences vs. Document Clustering

The key problem is avoiding False Positives during automatic annotation of articles with ontology terms, as, for example, in the automatic annotation of PubMed articles with GO or MeSH terms performed by GoPubMed (Doms and Schroeder, 2005).

One key idea is that term co-occurrences in a training dataset of manually annotated articles (e.g., GOA) can help to solve the problem. With the co-occurrence graph, we can define term groups which are associated with a given term. For example, does 'cell proliferation' co–occur frequently with 'development'? This knowledge can be utilized in two ways:

- If a document under examination contains 'development' but none of its co-occurring terms from the hand-curated data, then 'development' is likely to be a False Positive.

- If a document does not contain 'development', but some of its frequently co-occurring terms, then it is likely to be a False Negative (the document is not automatically annotated as developmental because 'development' does not literally appear, but actually talks about development).

Co-occurrence of terms can be defined in various ways. Here, we examine two approaches:

- First, we calculate the likelihood of *co-occurrence*, i.e., the number of documents in which two terms co-occur divided by the total number of documents. This likelihood does not take into account the probability of each of the terms occurring. A rarely occurring term should get a higher score than a frequently occurring term. Therefore, we define a score based on the BLOSUM (Henikoff and Henikoff, 1992) approach to substitution matrices in sequence comparison. In this context, this $log - odds$ score, TC_{score}, is the logarithm of the probability of two terms co-occurring divided by the probability of the two terms occurring.

- Besides the use of term co-occurrences, we also *cluster* documents by automatically derived annotations of the GoPubMed algorithm (Doms and Schroeder, 2005). For clustering, we use MULIC, a clustering algorithm for categorical data (Andreopoulos *et al.*, 2007). The clusters are organized in layers and for each layer of documents we assign an annotation based on the likelihood method in the first step. In the same way that co-occurring terms can give a clue for the correct annotation, grouping documents with similar annotations can further improve precision and recall.

[4]For details see Section 3.3.2, Datasets. To access the corpora, see Appendix A

development occurrence			
	GoPubMed	**GOA**	**Number of documents**
False Negative (FN)	–	✓	122
True Positive (TP)	✓	✓	109
False Positive (FP)	✓	–	100

Tab. 3.1: Benchmark dataset for 'development' based on GOA.

3.2.1 Datasets

Our training and test datasets are GOA and GoPubMed, respectively. GoPubMed represents articles *automatically* annotated with GO terms (Doms and Schroeder, 2005), while GOA represents articles *manually* annotated with GO terms (Camon *et al.*, 2005, 2004). GoPubMed consists of approximately 15,000,000 articles (PubMed) and GOA consists of approximately 34,000 articles. We map each GoPubMed article's annotations onto the corresponding subsection of the GOA corpus.

We manually generated three datasets containing True Positives (TP), False Positives (FP) and False Negatives (FN), with respect to the 'development' annotation (see Table 3.1). Then, we united these datasets into one dataset of 331 articles in total.

GoPubMed and GOA Statistics

The key data for the first step of our approach are terms co-occurring with 'development'. Tables 3.2 and 3.3 show the top ten terms associated with 'development' according to the number of co-occurrences and the TC_{score}, respectively. Both tables are broken down into terms according to GoPubMed's automated, comprehensive, but more error prone annotation and GOA's manual, less comprehensive, but higher quality annotation. The first row in Table 3.2 shows that 'cell' is the term appearing most frequently with 'development' in GoPubMed. The relatively low TC_{score} (negative unlikely, positive likely) reflects that the term is very general and hence not very predictive for 'development'. However, 'cell' is very effective for separating articles on cell biology from medical abstracts. The most frequently co-occurring term in GOA is 'cell proliferation'. It also has a good TC_{score}.

Table 3.3 shows the top terms according to the TC_{score}. The first line shows for GoPubMed 'petal development', which is clearly related to 'development' as it is a more specific term in the ontology, while the GOA annotation shows the extremely specific and term '3-mercaptopyruvate sulfurtransferase activity', which co-occurs only once. As GOA is limited in size, high TC_{scores} come with low co-occurrences. The reason could be that very specific proteins like TBX4, YY1, etc. are indicative of correct annotation with 'development' and these proteins are in turn correlating very well to the very detailed ontology terms listed in table 3.3.

The co-occurrence of terms can be extended to pairs frequently co-occurring with 'development'. Tables 3.4 and 3.5 summarize the joint probabilities of 'development' with the two most frequent terms. For GoPubMed (Table 3.4) the terms cover cell growth and differentiation in general, while for GOA (Table 3.5) the terms related to transcription are prominent. Both tables appear intuitively meaningful topics to indicate abstracts on cell biology.

3.2.2 Methodology

Our objective is to find two classes of articles, those that: *a.* should *not* be annotated with 'development', i.e., False Positives (FPs), and *b.* should be annotated with 'development', i.e., False Negatives (FNs) or True Positives (TPs). We map the annotations in each GoPubMed article to a graph representing co-occurrences of annotations in $\sim 34,000$ *manually* annotated articles in GOA. Based on several proba-

GoPubMed			GOA		
Term name	**cooc.**	TC_{score}	**Term name**	**cooc.**	TC_{score}
cell	200142	0.21	cell proliferation	25	2.40
growth	80751	0.62	transcription factor activity	23	1.29
biosynthesis	69146	0.17	regulation of transcription, DNA - dependent	22	1.95
cell development	46722	2.56	protein binding	20	-0.21
viral life cycle	45527	-0.01	nucleus	20	-0.04
antigen binding	45448	0.06	signal transduction	17	0.9
brain development	39119	0.22	integral to plasma membrane	15	0.66
cellularization	35330	0.4	DNA binding	14	0.8
binding	35042	-0.14	cytoplasm	11	-0.21
regulation of biological process	33777	0.45	apoptosis	11	1.88
behavior	33306	0.099	immune response	10	1.25

Tab. 3.2: The top 10 GO annotations in GoPubMed and GOA, according to their co-occurrence with 'development'.

bilistic metrics described below we infer the likelihood that 'development' should or should not annotate each article.

Overview

Our methodology uses a co-occurrence graph based on manually annotated GOA articles. We find co-occurring terms in all GOA articles and build a co-occurrence graph representing how frequently pairs of GOA terms co-occur in all GOA articles. The nodes represent annotations and edges represent the frequency of co-occurrence of two annotations. We view each GoPubMed article as representing co-occurring GoPubMed annotations. Our approach involves mapping each GoPubMed article onto the co-occurrence graph of manual GOA annotations. Each GoPubMed article is mapped to the nodes and edges of the GOA co-occurrence graph. Then, we use several metrics to estimate the likelihood of a 'development' annotation being appropriate for the GoPubMed article, based on an n-word of n annotations that are neighbors of 'development' in the GOA co-occurrence graph.

GOA is sparsely annotated because of the effort required in assigning manual annotations. For this reason, we use the GOA co-occurrence graph such that high correlations of annotations with 'development' are considered more significant than low correlations. In the GOA co-occurrence graph an annotation a_i's low correlation with 'development' is not a very strong sign for a FP. On the other hand, an annotation a_i's high correlation with 'development' is a stronger sign for a FN or TP. With this rationale, we assign to each article a 2-word, including 'development' and the article's annotation most closely correlated with 'development' in the GOA co-occurrence graph. We use these 2-words with probabilistic metrics to assess which articles are most likely to be relevant to 'development' (TPs/FNs); the rest of the articles are considered more likely not to be relevant to 'development' (FPs). The use of 2-words is specific to our application, which classifies articles as TPs/FNs based on the annotation most correlated with 'development'; however, n-words for any n could potentially be used.

We also propose a clustering methodology for finding groups of GoPubMed articles (clusters or subclusters) that are FPs or FNs/TPs. Our clustering methodology improves the results, since many GoPubMed articles are incomplete with missing annotations (FNs) or have wrong annotations that should be filtered out (FPs). Moreover, most annotations occur infrequently in GOA. Clustering allows

GoPubMed			GOA		
Term name	TC_{score}	**cooc.**	**Term name**	TC_{score}	**cooc.**
petal development	2.55	78	3 - mercaptopyruvate sulfurtransferase activity	4.95	1
sepal development	2.55	19	hydrolase activity, acting on acid anhydrides, catalyzing transmembrane movement of substances	4.95	1
stamen development	2.55	80	intramolecular transferase activity, phosphotransferases	4.95	1
carpel morphogenesis	2.55	3	carbon-nitrogen ligase activity, with glutamine as amido-N-donor	4.95	1
sepal morphogenesis	2.55	2	lipoate-protein ligase B activity	4.95	2
stamen morphogenesis	2.55	2	transcription initiation factor activity	4.95	2
carpel structural organization	2.55	1	sigma factor activity	4.95	1
establishment of petal orientation	2.55	2	glutamyl-tRNA(Gln) amidotransferase activity	4.95	1
meristem development	2.55	215	protein prenylation	4.95	1
gut development	2.55	578	protein amino acid prenylation	4.95	2
regulation of post-embryonic development	2.55	13	alkane 1-monooxygenase activity	4.95	1

Tab. 3.3: The top 10 GO annotations in GoPubMed and GOA, according to their TC_{score} with 'development'.

GoPubMed		
Term name A	**Term name B**	**Prob(A,B,development)**
cell growth	growth	10^{-3}
cell	cell growth	10^{-3}
cell	cell surface	10^{-3}
cell	cell differentiation	1.2×10^{-3}
cell	regulation of biological process	1.3×10^{-3}
cell	binding	1.3×10^{-3}
cell	cellularization	1.7×10^{-3}
cell	antigen binding	1.7×10^{-3}
cell	biosynthesis	2.4×10^{-3}
cell	growth	2.7×10^{-3}
cell	cell development	3.1×10^{-3}

Tab. 3.4: The top 10 pairs of non-'development' GO annotations in GoPubMed, according to their probability of co-occurring with 'development'.

GOA		
Term name A	**Term name B**	**Prob(A,B,development)**
signal transduction	cell proliferation	1.5×10^{-4}
DNA binding	transcription factor activity	1.7×10^{-4}
DNA binding	nucleus	1.7×10^{-4}
transcription factor activity	transcription from RNA polymerase II promoter	1.7×10^{-4}
protein binding	nucleus	1.7×10^{-4}
protein binding	regulation of transcription, DNA - dependent	1.7×10^{-4}
nucleus	regulation of transcription, DNA - dependent	1.7×10^{-4}
transcription factor activity	regulation of transcription, DNA - dependent	2×10^{-4}
protein binding	cytoplasm	2×10^{-4}
transcription factor activity	nucleus	2.6×10^{-4}

Tab. 3.5: The top 10 pairs of Non-'development' GO annotations in GOA, according to their probability of co-occurring with 'development'.

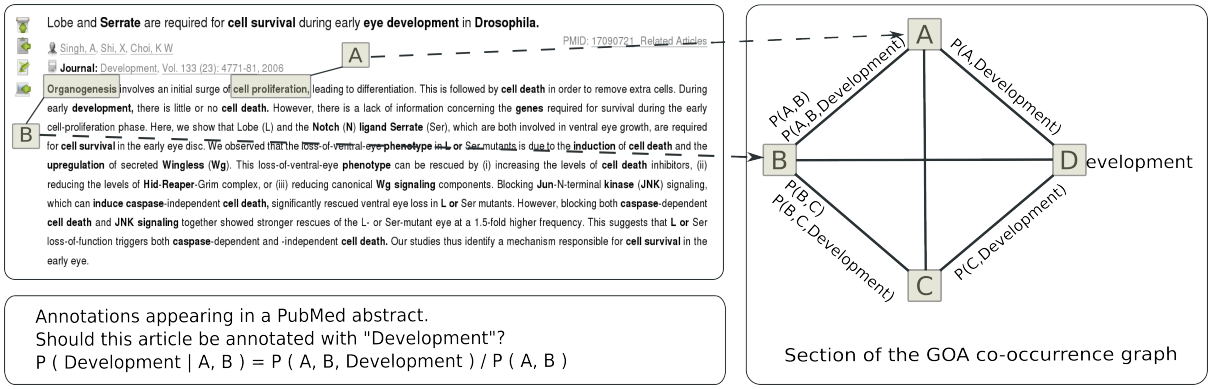

Fig. 3.1: Mapping a GoPubMed article's annotations onto the GOA co-occurrence graph. A GoPubMed article points to the edges of the GOA co-occurrence graph corresponding to pairs of co-occurring GOA annotations.

to aggregate information on the occurrences of annotations over all GoPubMed articles. Clustering allows to build groups of GoPubMed articles and to identify the union set of articles' annotations in a group. We can map the union of annotations in a group of GoPubMed articles onto the GOA co-occurrence graph, rather than each individual article's annotations. This allows to return to the user for examination groups of articles that are likely to be FPs or FNs/TPs, rather than individual articles. This makes the process of looking for FPs and FNs/TPs more accurate. This also makes the process faster by avoiding redundant mappings of GoPubMed articles with the same annotations to the GOA co-occurrence graph.

Co-occurrence Graph of GOA Annotations

In order to formalize the notion of GO annotations' co-occurrences, we consider pairs of GO terms that appear in the same article's abstract and we represent all such pairs of GO terms in a GOA co-occurrence graph. We created a GOA co-occurrence graph for the set of all manually annotated articles in the GOA database. GOA articles annotated with 'development' are most likely to be TPs. Each node represents a GO annotation. An edge between nodes a_1 and a_2 represents:

- A probability $P(a_1, a_2, development)$ representing the likelihood of co-occurrence of the terms a_1, a_2 and 'development' in all GOA articles. For an edge between a_1 and the 'development' node this is equal to $P(a_1, development)$.

- A real number, called a TC_{score}, representing the frequency $TC_{score}(a_1, a_2)$ of the terms' a_1, a_2 co-occurrence over all articles. The TC_{score} for a pair a_1 and a_2 is estimated as: $log\frac{P(a_1,a_2)}{P(a_1)P(a_2)}$.

Figure 3.1 shows an example of mapping a set of GoPubMed annotations for an article onto the GOA co-occurrence graph.

Using special metrics we assess the likelihood of 'development' being a TP/FN or FP for an article.

First TC_{score} metric for finding 'development' TPs, FPs, FNs

The first metric is the TC_{score}, which was presented in the previous section. For an article we find the annotation a_1 which is the most correlated to 'development' in the GOA co-occurrence graph. (If there is a tie, it will not affect the result). Then, we assign to the article the $TC_{score}(a_1, development)$.

The rationale for considering only one non-'development' annotation a_1 for each article, as described earlier, is that we consider annotations most closely correlated with 'development' as most reliable for classifying a GoPubMed article as FN/TP.

Second probabilistic metric for finding 'development' TPs, FPs, FNs

We are given a set of annotations $a_1, \cdots, a_n$ (for our purposes n=1) that are prominent in an article (or a group of articles derived via clustering). Suppose the GOA co-occurrence graph suggests that given these annotations $a_1, \cdots, a_n$ co-occurring in an article, the 'development' annotation has a probability π of correctly describing this article (TP/FN). Then, we estimate the likelihood that 'development' is a correct annotation for a GoPubMed article, based on several manual GOA annotations $a_1, \cdots, a_n$. This pseudo-Bayesian application is a simplification of strict statistical Bayesian rules for making our method practically useful. The rationale behind this application stems from the following motivating arguments: *a*. The relatively infrequent automatic GoPubMed annotations $a_1, \cdots, a_n$, are usually less likely to be FNs or FPs than 'development'. *b*. GOA manual annotations $a_1, \cdots, a_n$ are less likely to be FNs or FPs than automatic GoPubMed annotations.

We find FNs/TPs by initially looking for GoPubMed articles with annotations most frequently co-occurring with 'development' in GOA. We rank as most likely FNs/TPs the GoPubMed articles with annotations most frequently co-occurring with 'development' in GOA.

The GoPubMed articles ranked as most likely FPs are those with annotations that co-occur less frequently with 'development' in GOA.

For estimating the likelihood of co-occurrence, we use the following probability:

$$P(development, a_1, \cdots, a_n) = P(development|a_1, \cdots, a_n)P(a_1, \cdots, a_n) =$$

$$\frac{P(a_1, \cdots, a_n|development)P(development)P(a_1, \cdots, a_n)}{P(a_1, \cdots, a_n)}$$

$$where\ \{a_1, \cdots, a_n\} = set\ of\ n\ GoPubMed\ annotations\ that$$

$$co-occur\ in\ GOA\ co-occurrence\ graph\ with\ 'development'$$

In our case, we are only interested in estimating the part $P(a_1, \cdots, a_n|development)$, since $P(development)$ remains constant and it will not affect our decision on which articles should or should not be annotated with 'development'.

We consider one annotation a_1 for each GoPubMed article, the one most closely correlated with 'development' in the GOA co-occurrence graph. Thus, we just need to estimate:

$$P(a_1|development) = \frac{P(a_1, development)}{P(development)}$$

For fast retrieval of the values of this measure we use the co-occurrence graph, as previously described. This is one of the purposes of our co-occurrence graph, to serve as a data structure allowing for quick retrieval of the probabilities.

Threshold for separating likely FNs/TPs from FPs

We set a threshold for each of the two metrics described above, to separate:

1. GoPubMed articles that are 'development' FNs or TPs. These articles are often manually annotated as 'development' in GOA.
2. GoPubMed articles that are 'development' FPs. These articles are automatically annotated as 'development' in GoPubMed, but often do not have this manual annotation in GOA.

By comparing each GoPubMed article's annotations to the GOA co-occurrence graph we establish a *threshold* for the values of each metric previously described. The threshold separates articles into two groups: FNs/TPs from FPs. We examine the appropriate value of *threshold* in the next section on experiments.

Input: a set S of articles with GoPubMed annotations;
Parameters: (1) $\delta\phi$: the increment for ϕ;
(2) *threshold* for ϕ : the maximum number of annotations that can differ between an article and the mode of its cluster;
Default parameter values: (1) $\delta\phi = 1$;
(2) *threshold* = the number of distinct annotations in GoPubMed;
Output: a set of clusters;
Method:

1. Order the articles from lowest to highest degree;
2. Insert the first article into a new cluster, use the article as the mode of the cluster, and remove the article from S;
3. Initialize ϕ to 1;
4. Loop through the following until S is empty or $\phi >$ *threshold*
  a. For each article o in S
    i. Find o's closest cluster c by using the similarity metric to compare o with the modes of all existing cluster(s);
    ii. If the number of different annotations between o and c's mode is larger than ϕ, insert o into a new cluster
    iii. Otherwise, insert o into c and update c's mode;
    iv. Remove article o from S;
  b. For each cluster c, if there is only one article in c, remove c and put the article back in S;
  c. If in this iteration no articles were inserted in a cluster with $size > 1$, increment ϕ by $\delta\phi$.

Fig. 3.2: The MULIC clustering algorithm (Andreopoulos *et al.*, 2007).

Clustering

We use the MULIC clustering algorithm for partitioning the articles into groups of articles with similar GoPubMed annotations. Consider a group (cluster or subcluster) of articles. If 'development' is a TP or FN for the group of articles, then, the group likely contains a set of GoPubMed annotations which in the GOA co-occurrence graph are correlated with one another and 'development'. If the group's set of GoPubMed annotations are not connected to 'development' in the GOA co-occurrence graph, then 'development' is more likely to be a FP for the group of articles.

The objects to be clustered are the GoPubMed articles. Clustering decisions are based on each article's set of GoPubMed annotations. MULIC clusters consist of layers, where each layer corresponds to a different value of the similarity criterion used for inserting articles in clusters.

Each MULIC cluster has a *mode*, which is the union of annotations of all article members of the cluster.

MULIC ensures that, when each article o is clustered, it is inserted into the cluster c with the most similar mode μ_c, thus, maximizing the similarity between article and mode. The similarity metric is defined as follows:

$$similarity(o, \mu_c) = |o \cap \mu_c|$$

where o is an article in the dataset and μ_c is the mode of the cluster c in which o is to be inserted.

Figure 3.2 shows the MULIC clustering algorithm, as used in our application. The algorithm starts by reading all articles from the input file and storing them in S. Objects (articles) are ordered from lowest to highest degree, where the degree is the number of annotations. The first article is inserted in a new cluster, the article becomes the mode of the cluster and the article is removed from S. Then, it continues iterating over all articles that have not been assigned to clusters yet, to find the closest cluster. In all iterations, the closest cluster for each unclassified article is the cluster with the highest similarity between the cluster's mode and the article, as computed by the similarity metric.

The variable ϕ is maintained to indicate how high the dissimilarity is allowed to be between an article

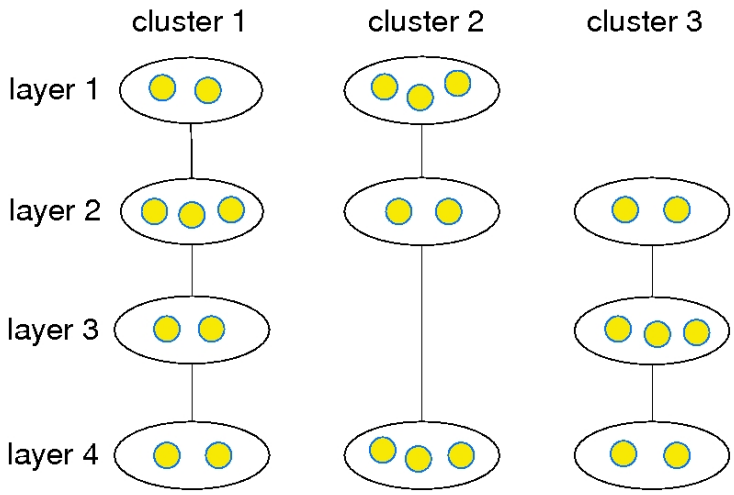

Fig. 3.3: A MULIC cluster consists of one or more layers representing dissimilarities between the articles and mode. Ovals are layers and circles are articles.

and the closest cluster's mode, for the article to be inserted in the cluster. The dissimilarity metric is defined as follows:

$$dissimilarity(o, \mu_c) = |(o - \mu_c) \cup (\mu_c - o)|$$

Initially ϕ equals 1, meaning that only one annotation can differ between an article and the closest cluster's mode. If the number of different annotations between the article and the closest cluster's mode is greater than ϕ, then, the article is inserted in a new cluster on its own, else, the article is inserted in the closest cluster and the mode is updated.

At the end of each iteration, all articles assigned to clusters of size one have their clusters removed so that the articles will be re-clustered at the next iteration. This ensures that the clusters that persist through the process are only those containing at least two articles. Articles assigned to clusters of size greater than one are removed from the set of unclassified articles S, so those articles will not be re-clustered.

At the end of each iteration, if no articles have been inserted in clusters of size greater than one, then the variable ϕ is incremented by $\delta\phi$. Thus, at the next iteration the criterion for inserting articles in clusters will be more flexible. The iterative process stops when all articles are classified in clusters of size greater than one, or ϕ exceeds a user-specified *threshold*. If the *threshold* equals its default value, the process stops when all articles are assigned to clusters of size greater than one.

The MULIC algorithm can eventually classify all articles in clusters, even if the closest cluster to an article is very dissimilar, because ϕ can continue to increase until all articles are classified. Even in the extreme case, where an article o has only zero or one annotation similar to the mode of the closest cluster, it can still be classified when ϕ reaches a high value.

Figure 3.3 illustrates what the results of MULIC look like. Each cluster consists of one or more different "layers". The layer of an article represents how high the article's dissimilarity was to the mode of the cluster when the article was assigned to the cluster. The cluster's layer in which an article is inserted depends on the value of ϕ. Bottom layers such as 1000 correspond to higher values of ϕ and have a lower coherence - meaning a higher average dissimilarity between all pairs of articles in the layer. MULIC starts by inserting as many articles as possible in top layers - such as layer 1 - and then moves to bottom layers, creating them as ϕ increases.

If an unclassified article has equal similarity to the modes of the two or more closest clusters, then the algorithm tries to resolve this 'tie' by comparing the article to the mode of the top layer of each of these clusters - the top layer of a cluster may be layer 1 or 2 and so on. Each cluster's top layer's mode was stored by MULIC when the cluster was created, so it does not need to be recomputed. If the article

has equal similarity to the modes of the top layer of all of its closest clusters, the article is assigned to the cluster with the highest bottom layer. If all clusters have the same bottom layer then the article is assigned to the first cluster, since there is insufficient data for selecting the best cluster.

Ordering the Articles before Clustering

When running MULIC with different random orderings of the dataset objects (articles), the result is often different. The modes and clusters are influenced most by the annotations of the articles that are clustered first in top cluster layers. It makes more sense to cluster first the articles of low degree (i.e., articles with few annotations) and last the articles of high degree.

Two articles of high degree are unlikely to have the exact same annotations, thus, it is unlikely that there will be many articles of high degree in top cluster layers. By ordering the articles and presenting them to the clustering process from low to high degree, and by gradually relaxing ϕ, the clusters get an onion-layered structure where articles in top layers have similar sets of annotations and articles in bottom layers have less similar sets of annotations.

Characteristics for GoPubMed Article Clustering

We implemented several characteristics specific for GoPubMed article clustering, such as the mode's updating and the dis/similarity metric as described in the previous section.

While the MULIC clustering algorithm follows the basic framework of k-Modes (Huang, 1998), it has substantially different characteristics. *First*, clusters are layered. *Second*, the number of clusters is not specified by the user - clusters are created, removed or merged, as the need arises. K-Modes requires the user to specify the number of clusters and the algorithm builds and refines the specified number of clusters. *Third*, all MULIC clusters are of size two or greater.

3.2.3 Experimental evaluation

Each article had an original classification as FP, FN or TP with respect to the 'development' annotation. Our goal was to find out whether an article could be classified correctly as FP, FN, or TP based on its mapping to the GOA co-occurrence graph. We are interested in articles that were erroneously automatically annotated as 'development' (FPs), or should be automatically annotated as 'development' (FNs or TPs).

In order to evaluate the success of our methodology for separating likely FPs from FNs and TPs we used the precision/recall measure, as described next.

Precision (P) and Recall (R)

Without loss of generality assume that the optimal mapping assigns class c_i to the retrieved group of articles g_i. There are two known classes in our test dataset S: *a.* c_1 consisting of articles known to be FNs/TPs and *b.* c_2 consisting of articles known to be FPs. Our result consists of two groups of articles, g_1 and g_2, the former believed to be FN/TP articles and the latter believed to be FP articles. We define precision, P_i, and recall, R_i, for a group of articles g_i, $1 \leq i \leq 2$ as follows (Andritsos *et al.*, 2004):

$$P_i = \frac{|g_i \cap c_i|}{|g_i|} \quad and \quad R_i = \frac{|g_i \cap c_i|}{|c_i|}$$

P_i and R_i take values between 0 and 1 and, intuitively, P_i measures the accuracy with which group g_i reproduces class c_i, while R_i measures the completeness with which group g_i reproduces class c_i. We define the precision and recall of the result as the weighted average of the precision and recall of each

group of articles. More precisely:

$$P = \sum_{i=1}^{k} \frac{|c_i|}{|S|} P_i \;\; and \;\; R = \sum_{i=1}^{k} \frac{|c_i|}{|S|} R_i$$

We think of precision, recall, and misclassification as indicative values (percentages) of the ability of our methodology to reconstruct the existing classes in the dataset.

Results for Classifying Articles Individually

In order to classify articles we define a *threshold* drawing a line that separates likely FPs from FNs and TPs. Tables 3.6 and 3.7 show the precision and recall achieved for the two metrics and different values of threshold. The precision and recall show how effectively each metric and threshold partition the articles into the classes of FPs and FNs/TPs.

As shown, for the first TC_{score} metric the best partitioning is achieved with a threshold value of 0. This points to the significance of the results, since 0 would be the natural choice for the TC_{score} threshold value for separating FPs from FNs/TPs.

For the second probabilistic metric (range from 0 to 1) the best partitioning is achieved with a threshold value of 0.00004. The best precision and recall for the second probabilistic metric are 0.77, slightly better than the first TC_{score} metric. The reason for the improved result may be that the first TC_{score} metric is slightly biased by considering in its denominator the likelihood of individual annotations' occurrences. While the second probabilistic metric just considers the likelihood of co-occurrences of annotations.

Results with Clustering

We clustered all 331 automatically annotated GoPubMed articles in our dataset. We used MULIC with its default parameter values. For clustering we excluded the GoPubMed 'development' annotations. We did not use the manually annotated GOA articles. We got 22 clusters, where each cluster had on average 4 layers. Most clusters had a top layer of 0 or 1, containing annotation sets that are representative of corpus groups of articles.

We consider each cluster as a distinct group of GoPubMed articles, the combined annotation set of which is mapped onto the GOA co-occurrence graph. Then, we classify each cluster as FP or FN/TP. Articles in different clusters might have dissimilar GoPubMed annotations and there is a large number of annotations in the dataset.

Then, we used the metrics previously described for finding whether the 'development' annotation is more likely to be a FP or FN/TP for a group of GoPubMed articles. We examined how accurately the neighborhood of the GOA co-occurrence graph corresponding to the group's articles' annotations reflects whether 'development' is or is not appropriate for the group.

Tables 3.8 and 3.9 show that the results with clustering are improved and the best precision reached is 0.82. The best threshold values are the same as without clustering. For the first TC_{score} metric the best partitioning is achieved with a threshold value of 0, while for the second probabilistic metric with a threshold value of 0.00004.

The main reason for the improved results with clustering is that we aggregate information on annotations of clusters of related GoPubMed articles. This way, articles that are incomplete with missing annotations have a less negative effect on finding 'development' FPs, FNs and TPs.

3.2.4 Conclusion

We have proposed and evaluated an approach for improving the quality of automatically annotated articles. This approach is based on co-occurrence graphs, which in our case we built on the basis of

Threshold	Precision	Recall
-1.0	0.74	0.74
-0.5	0.74	0.74
0	0.74	0.74
0.5	0.74	0.74
1.0	0.73	0.73
1.5	0.72	0.71
2.0	0.71	0.7
3.0	0.68	0.55

Tab. 3.6: Precision and Recall for the first TC_{score} metric and different *threshold* values (without MULIC clustering of articles).

Threshold	Precision	Recall
0	0.74	0.74
1×10^{-5}	0.74	0.74
2×10^{-5}	0.74	0.74
3×10^{-5}	0.77	0.77
4×10^{-5}	0.77	0.77
5×10^{-5}	0.77	0.77
6×10^{-5}	0.75	0.65
7×10^{-5}	0.75	0.65
8×10^{-5}	0.75	0.65
9×10^{-5}	0.77	0.61

Tab. 3.7: Precision and Recall for the second probabilistic metric and different *threshold* values (without MULIC clustering of articles).

Threshold	Precision	Recall
-1.0	0.82	0.72
-0.5	0.82	0.72
0	0.82	0.72
0.5	0.82	0.72
1.0	0.79	0.73
1.5	0.78	0.74
2.0	0.79	0.75
3.0	0.71	0.7

Tab. 3.8: Precision and Recall with MULIC clustering of articles, for the first TC_{score} metric and different *threshold* values.

Threshold	Precision	Recall
0	0.82	0.72
1×10^{-5}	0.82	0.72
2×10^{-5}	0.82	0.72
3×10^{-5}	0.82	0.75
4×10^{-5}	0.82	0.75
5×10^{-5}	0.82	0.75
6×10^{-5}	0.77	0.75
7×10^{-5}	0.77	0.75
8×10^{-5}	0.77	0.75
9×10^{-5}	0.75	0.73
10^{-4}	0.75	0.73

Tab. 3.9: Precision and Recall with MULIC clustering of articles, for the second probabilistic metric and different *threshold* values.

GOA.

Even though we have focused on the 'development' annotation, the method proposed can be applicable to diverse annotations that are often FPs or FNs, such as 'cell' or 'growth' or 'determination of affect'.

One problem with this approach is that GOA is sparsely annotated because of the difficulty and effort required for manual annotations. For example, in GOA only 243 articles have a 'development' annotation. This may raise questions as to the statistical significance of mapping an automatically annotated GoPubMed article onto the GOA co-occurrence graph. Even though the GOA corpus that we used for the co-occurrence graph contains only 243 articles annotated with 'development', our results are still shown to be meaningful. As the GOA corpus increases, the statistical significance of the relationships in the co-occurrence graph will become stronger. Then, future experimental results will be even more meaningful for predicting FPs and FNs.

We further extend the co-occurrence approach (called **'Term Cooc'**) in the next section in the following ways:

1. We additionally disambiguate MeSH terms and use a larger training corpus to get the co-occurrence scores, since there exist $\sim 16,400,000$ documents to which experts have assigned MeSH terms.

2. We make use of the hierarchy structure in both GO and MeSH (given an ambiguous term α, the co-occurrence of α with a term β should not be lower than α's co-occurrence with any of β's descendants, 'Inferred Cooc'), whereas before we used term co-occurrence without any inference. We therefore investigate how two different hierarchies influence the performance of disambiguation.

3. We combine our graph–based decision function with a support vector machine, arranged in a co–training scheme, to learn and improve models without any labelled data.

4. We test the disambiguation performance in new larger benchmark datasets of varying curation quality.

3.3 Term Cooc vs. Closest Sense vs. MetaData

In this section we further extend the co-occurrence approach presented previously and develop two more approaches to word sense disambiguation, which use ontologies and metadata. The **'Term Cooc'** method defines a log-odds ratio for co-occurring terms including co-occurrences inferred from the ontology structure. The **'Closest Sense'** method assumes that the ontology defines multiple senses of the term. It computes the shortest path of co-occurring terms in the document to one of these senses. The **'MetaData'** approach trains a classifier on metadata. It does not require any ontology, but requires training data, which the other methods do not.

To evaluate these approaches we define a manually curated training corpus of 2600 documents for seven ambiguous terms from the Gene Ontology and MeSH (see Appendix A).

3.3.1 Methods

Terminology and classification of approaches

The types of relations between terms in the Gene Ontology (GO), the Medical Subject Headings (MeSH) and the Unified Medical Language System (UMLS) semantic network make them completely different knowledge sources (Bodenreider and Stevens, 2006). GO has a simple structure in the form of a directed acyclic graph and GO terms are interconnected via *is_a* and *part_of* relations. The semantics of relations used in MeSH make it a terminology rather than an ontology. Terms in MeSH are related through *A narrower than B* relations, giving users who are interested in *B*s the option to look at *A*s. The UMLS is considered to be a terminology integration system comprising over 150 biomedical vocabularies and relations like subClassOf or subPropertyOf between terms. Therefore, it is located in the space between a structured terminology and an ontology. The most popular semantic web formalisms for representing taxonomies, ontologies and terminologies in general are the Resource Description Framework (RDF) and the Web Ontology Language (OWL), with OWL being more suitable for ontologies and RDF sufficient for terminologies. The Simple Knowledge Organization Systems[5] (SKOS) is an area of work developing specifications and standards to support the use of knowledge organisation systems (KOS) such as thesauri, classification schemes, subject heading systems and taxonomies within the framework of the Semantic Web. SKOS provides a standard way to represent knowledge organisation systems using RDF. Lately, there have also been provided OWL translations of GO and MeSH by the responsible consortia.

We designed, implemented, and evaluated three WSD methods that we refer to as: *Closest Sense* **(CS)**, *Term Cooc* **(TC)**, and *MetaData* **(MD)**. Their differences are explained in the following:

Background knowledge: Closest Sense (CS) uses the UMLS semantic network; it represents an abstract as a list of UMLS terms occurring in the abstract. Term Cooc (TC) uses co-occurrences of terms in GO and MeSH, built from a curated dataset; it represents a document abstract as a list of GO and MeSH terms occurring in the abstract. The MetaData method (MD) uses metadata about the journal and title; it represents a document abstract as n-tuples of word stems and metadata.

Classification: Closest Sense uses shortest semantic distance of co-occurrences to sense. Term Cooc uses Support Vector Machines and co-occurrences from a training dataset for finding boundary between senses. MetaData uses the maximum entropy to model the behavior of the co-occurrence of contextual words and metadata with the ambiguous term.

Figure 3.4 gives an overview of the disambiguation performed by the three methods. 'Thrush' can refer to a mouth disease (oral candidiasis) or to a songbird (e.g., thrush nightingale). The CS method examines what appears in the same sentence and/or paragraph (e.g., 'mouth diseases' or 'oral ulcer')

[5] See SKOS http://www.w3.org/2004/02/skos/

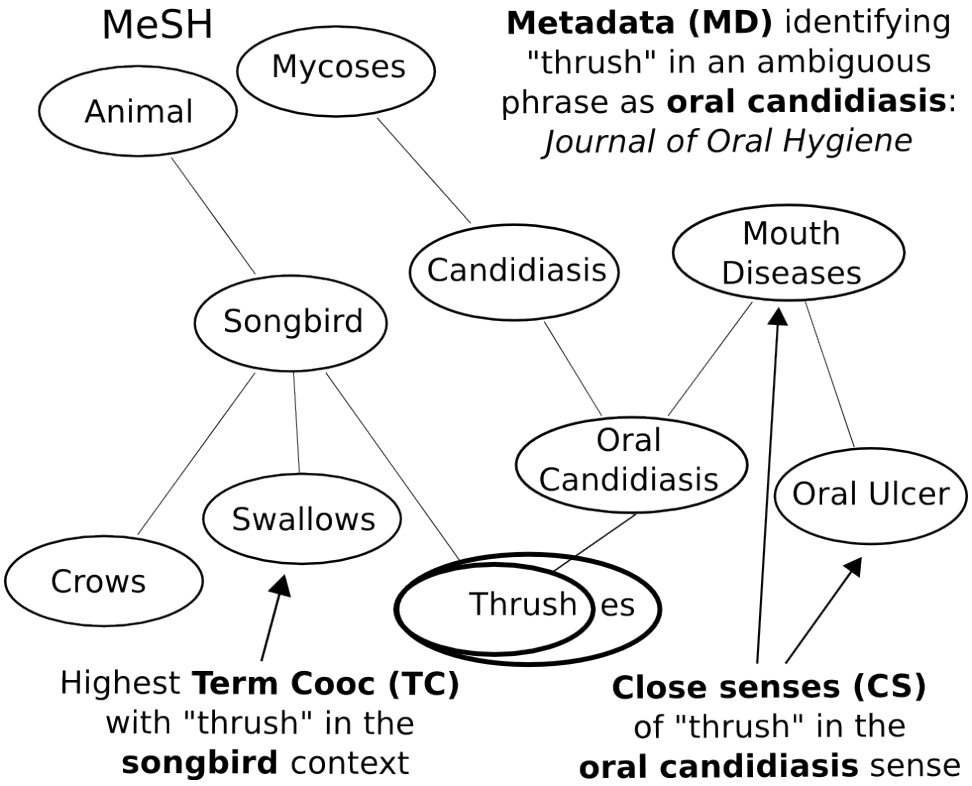

Fig. 3.4: Three disambiguation approaches for one term. *Thrush* is an ambiguous term, as its senses include *songbird* or *oral candidiasis*. This figure shows the possibilities for disambiguating 'thrush'. Solid edges are **is_a** relationships.

and then computes a similarity based on semantic distances to 'songbird' and 'oral candidiasis' in the UMLS semantic network, with the highest similarity determining the result. The TC method examines what appears in the same abstract (e.g., 'swallows') and then considers all known co-occurrences between taxonomy terms in the training corpus. The value of the highest co-occurrence determines the result, e.g., 'swallows' would have relatively high co-occurrence with 'thrush' songbird. The MD method uses metadata for the document and then decides based on what was previously learned about this metadata from training examples. If, for example, the article comes from the Journal of Oral Hygiene, then it is more likely that 'thrush' refers to 'oral candidiasis'.

Closest Sense method (CS)

This WSD approach was initially used to address the ambiguity problems in the MeatAnnot system (Khelif *et al.*, 2007) and developed by colleagues from INRIA Sophia Antipolis, in the framework of the Sealife[6] project. The main idea of the approach is the following: given a set of different senses of the ambiguous term, the co-occurring terms in the same text and the hierarchy where they belong (including the different types of relations), decide which sense is true based on the (shortest) distance to the senses of the co-occuring terms.

To clarify this, we can have the following sentence as an example: *'I also tracked **lipid profiles**, **HBA1C**, **blood pressure**, **body mass index**, **hostility** and nicotine use'*. The term 'blood pressure' can have three senses, namely 'organism function', 'diagnostic procedure' and 'laboratory or test result' (see Table 3.10). The senses of the co-occurring terms are 'laboratory procedure' (lipid profile), 'gene or genome' (HBA1C), 'diagnostic procedure' (body mass index), 'mental process' (hostility) and 'organic chemical' (nicotine). The sense of 'diagnostic procedure' for blood pressure is in average closer to the

[6] http://www.biotec.tu-dresden.de/sealife/

	Organism Function	Diagnostic Procedure	Laboratory or Test Result
lipid profile	0.291	0.862	0.167
HBA1C	0.166	0.167	0.3
body mass index	0.29	1	0.169
hostility	0.98	0.28	0.168
Average	0.431	**0.549**	0.201

Tab. 3.10: Disambiguation of 'blood pressure' with the Closest Sense approach. The term 'blood pressure' can have three senses, namely 'organism function', 'diagnostic procedure' and 'laboratory or test result'. The sense of 'diagnostic procedure' for blood pressure is in average closer to the senses of the co-occurring terms than the other candidate senses (it has the highest average semantic similarity).

senses of the co-occurring terms than the other candidate senses. For the example in Figure 3.4, the CS method determines the meaning of 'thrush' by examining what appears in the same sentence and/or paragraph (e.g., 'mouth diseases' or 'oral ulcer') and then computing a similarity based on semantic distances to 'songbird' and 'oral candidiasis' in the UMLS hierarchy; the highest similarity determines the result. Intuitively, with semantic distances, two senses are close if there exists a possibility to use them in a concise annotation graph.

Algorithm The 'Closest Sense' algorithm takes as input: (*i*) the ambiguous term τ, (*ii*) the vector V_τ of different senses of τ, (*iii*) the vector VC_τ of senses found in the context (sentence and/or paragraph containing the ambiguous term τ), and (*iv*) the UMLS semantic network.

First, the disambiguator builds a vector VC_τ of senses describing the context of the ambiguous term τ. This vector includes the senses of terms that are neighbours of τ. Then, it computes the similarity between each sense in vectors V_τ and VC_τ.

The resulting similarity is the average of similarities between senses in the two vectors. Finally, the sense in V_τ that has the highest average similarity to VC_τ is proposed as the best for τ.

Semantic distances The distance metrics used to find the correct sense are the *subsumption distance* and the *subtype-aware signature distance*.
The *subsumption distance* is the length of the shortest path between two *senses* in the hierarchy of senses, where the length of an individual subsumption link gets exponentially smaller with the depth of the senses it links in the hierarchy.
The *subtype-aware signature distance* is the length of the shortest path between two *concepts/terms* through the graph formed by the property types with their range links and domain links. With this new semantic distance we merge signature and hierarchies graphs. The main idea is to find a path between two concepts/terms by using the ontology structure (subClassOf relations between terms, subPropertyOf relations between properties) and the signature of relations (domain and range). The subtype-aware signature consists of relations in the hierarchy (subClassOf, subPropertyOf) additional to the common signature (domain and range of a property). It is *aware* of the properties of a term (signature), the position of the term in the hierarchy (subClassOf relations) and the hierarchy of the properties (subPropertyOf relations). The formal definitions of the distance metrics are explained in Khelif *et al.* (2008).

Figure 3.5 provides an example of the *subtype-aware* signature distance calculation between two terms in the UMLS semantic network. 'Body_Part_Organ_or_Organ_Component' is a subClassOf 'Fully_Formed_Anatomical_Structure', which belongs to the signature of the relation 'produces'. This relation has as range 'Organic_Chemical' which is a superClassOf 'Amino_Acid_Peptide_or_Protein'.

The *optimized distance* is a combination of the *subsumption distance* and the *signature distance*, parameterized with three *optimal weights*, w_{sig} for signatures, $w_{subclass}$ for class subsumption links and $w_{subprop}$ for property subsumption links. From a first experiment on the UMLS WSD test collection

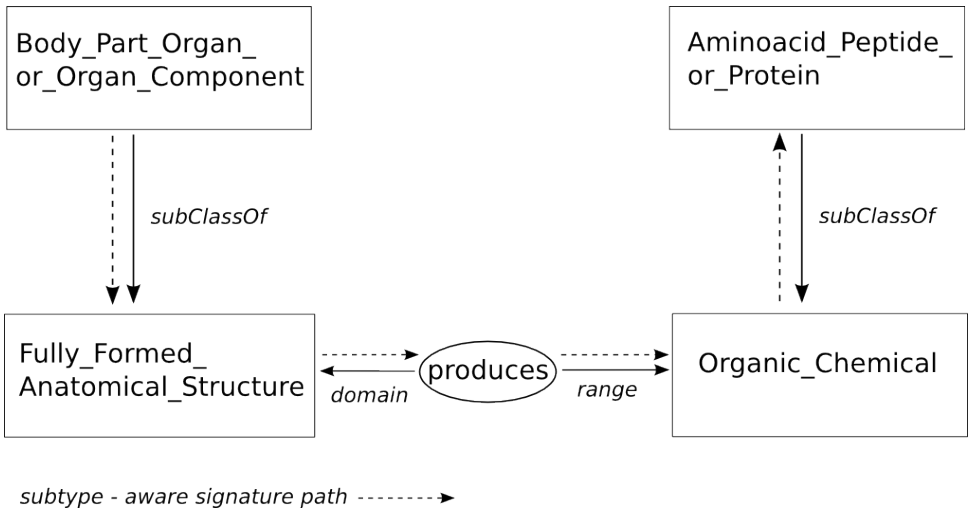

Fig. 3.5: Subtype-aware signature calculation. The figure shows the path between the UMLS terms 'Body_Part_Organ_or_Organ_Component' and 'Amino_Acid_Peptide_or_Protein'. The edges describe relations between entities (in our case, the subtype-aware-signature and its sub-properties) and nodes consist of classes and relations of the ontology. 'Body_Part_Organ_or_Organ_Component' is a subsumption of 'Fully_Formed_Anatomical_Structure', which belongs to the signature of the relation 'produces'. This relation has as range 'Organic_Chemical' which is a super-class of 'Amino_Acid_Peptide_or_Protein'. The length of this path is 4.

(Weeber *et al.*, 2001) where we tested different weights starting from a distance favouring the class subsumption relation to a distance favouring the signature relation, we ended up in the following optimal weights giving the best accuracy: $w_{sig} = 0.4$, $w_{subclass} = 0.2$ and $w_{subprop} = 0.4$.

Term Cooc method (TC)

The Term Cooc method relies on selecting the term that most frequently co-occurs with the ambiguous term in the training corpus. It selects the highest co-occurring term with the ambiguous term for defining the given sense as true or false. In order to formalize the notion of term co-occurrences (GO or MeSH), we consider pairs of GO/MeSH terms that appear in the same abstract and we represent all such pairs of terms in a manually annotated GOA or MeSH co-occurrence graph (see training with co-occurrence graphs subsection below). Each node in the co-occurrence graph represents a GO or MeSH manual annotation. An edge between nodes α and β represents a real number, the log-odds score, representing the frequency $log - odds(\alpha, \beta)$ of the terms' α and β co-occurrence over all articles, weighted by their total number of occurrences (TC_{score} previously defined in Section 3.2.2).

For the example in Figure 3.4, the TC method determines the meaning of 'thrush' by examining what appears in the same abstract (e.g., 'swallows') and then considering all known co-occurrences between ontology terms in a training corpus; the value of the highest co-occurrence determines the result, e.g., 'swallows' would have relatively high co-occurrence with 'thrush' songbird.

Algorithm First, we use a simple threshold considering how close to an ambiguous term the highest co-occurring term (of the ones in the article) is; if below a user-defined threshold θ, the ambiguous term is negative, else it is positive with respect to the term. Second, we use Support Vector Machines (SVMs) trained on all tokens of a text (Klinkenberg and Joachims, 2000).

The method first runs a binary SVM against a set of articles ordered by maximum co-occurrence with the ambiguous term. The highest and lowest 10% of articles in the set are labelled as positive and negative, respectively; then the SVM is trained on lower 10% (article with least co-occurring term with the ambiguous term) and upper 10% (article with highest co-occuring term with the ambiguous term). After the initial convergence is achieved, the error (wrongly classified vectors) will be low, likely near 0.

The algorithm next improves this result by iteratively re-classifying the remaining articles that have less extreme co-occurrences with the ambiguous term, one-by-one, followed by re-training the SVM on the newly relabelled data set. This continues until no more articles are left. The steps are:

1. *Set* S = Order articles based on their highest co-occurring GO/MeSH annotation (from the co-occurrence graph) with the ambiguous term.
2. T = lowest and highest 10% of S; label T as negative and positive; train SVM with T; remove T from S.
3. For $s \in S$: move s to T; classify s; re-train SVM with T.

Training with co–occurrence graphs These graphs are used in the TC method for *training*. For WSD of GO terms, the co-occurrence graph was derived from the Gene Ontology Annotations (GOA) (Camon *et al.*, 2004), and for the MeSH terms from the Medical Subject Headings (MeSH) (Nelson *et al.*, 2001). GOA represents articles manually annotated with GO terms and consists of $\sim 34,000$ articles. There exist $\sim 16,400,000$ documents to which experts have assigned MeSH terms. We found co-occurring terms in the GOA and MeSH annotated articles and built a co-occurrence graph representing how frequently pairs of GO or MeSH terms co-occur. Nodes represent annotations and edges represent the frequency of co-occurrence of two annotations in the same article, normalized based on each GOA/MeSH annotation's individual occurrence frequency in the specific corpus.

Training with inferred co–occurrences We extend the co-occurrences in a hierarchical fashion to ensure that given a GOA-derived co-occurrence between a pair of terms, $GOAcooc(\alpha, \beta)$, the ancestors of α and β in the ontology are updated with the co-occurrence such that only the maximum co-occurrence is kept. This is important given the few annotations in GOA and the is_a relationships between GO terms, since ancestors inherit the co-occurrences of their children.

With the inferred co-occurrences, given an ambiguous term α, the co-occurrence of α with a term β will not be lower than α's co-occurrence with any of β's descendants (see Figure 3.6).

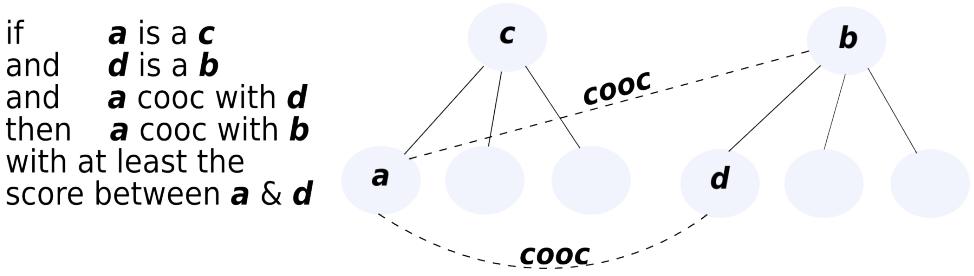

Fig. 3.6: Inferred co-occurrences (Inferred Cooc). Given an ambiguous term ***a***, the co-occurrence of ***a*** with a term ***b*** will not be lower than ***a***'s co-occurrence with any of ***b***'s descendants.

MetaData method (MD)

As an alternative method for WSD we use a maximum entropy approach as described in (Berger *et al.*, 1996; Pietra *et al.*, 1997). Maximum entropy models have been successfully used in tasks like part of speech tagging, sentence detection, prepositional phrase attachment, and named entity recognition.

For the example in Figure 3.4, the MD method[7] determines the meaning of 'thrush' by using metadata for the document and then deciding based on what was previously learned about this metadata from training examples. The metadata used are n-tuples of word stems from different scopes, namely the

[7]Implementation of the MetaData method by Dr. A. Doms (Bioinformatics group, BIOTEC, TU Dresden, http://www.biotec.tu-dresden.de/~adoms/)

paper title, the sentence including the ambiguous term or the whole abstract, the journal title as well as the publication period, since some topics can be popular in different decades. The occurrence of contextual words and phrases in a text together with a potentially ambiguous term can be seen as a random process. Maximum entropy modelling aims at modelling the behavior of this random process. Provided a large amount of training examples, the algorithm automatically extracts a set of relationships inherent in the examples, and then combines these rules into a model of the data that is both accurate and compact.

Algorithm The *training* and *test* data in our case are sentences containing the potentially ambiguous term flagged with the sense. Each training example, one sentence each, is represented as a set of features. An implementation of the Porter stemmer is used (Porter, 1980) and as features we select n-tuples of word stems and meta information of the document, such as the journal and title words and the publication period (10 years ranges).

The *implementation*[8] takes a series of events to train a model. Each event is a configuration of binary relations associated with a label. The resulting model is applied to an unknown configuration of binary relations. The result is the predicted probability for the previously trained outcomes. MeSH terms already assigned to the articles are excluded, for the performance evaluation to be independent of them.

Given the abstract of a scientific article and the ambiguous term, the steps followed are:

1. extract binary features (n-tuples of word stems from different scopes - title, sentence, entire abstract -, publication period, journal title)
2. get scalar product of feature vector and model (vector based on training)
3. the result is the probabilities for predefined outcomes (in this case True or False)
4. if above a threshold 0.5, the term is True, else False.

As an illustrating example of the features extracted, articles mentioning 'signal transduction', 'kinase', 'embryo', 'neuron' or 'stage' are more likely to refer to 'multicellular organismal development' than to another sense, such as development of an algorithm or a disease in an organism. Some extracted features indicating positively the sense of 'psychological inhibition' are the journals *'Physiol Behav'*, *'J Abnorm Psychol'* and phrases such as 'conditioned stimulus', 'emotion regulation', 'anxiety' and 'fear'. On the other hand, when an article contains mentions such as 'diabetes', 'pH', 'tumor', 'antibody', 'enzyme', 'protein' and 'membrane', then it is more likely to refer to other senses of the ambiguous term 'inhibition', such as e.g., 'enzyme inhibition'.

3.3.2 Experimental setup

Classification task and limitations

The disambiguation performed here is mainly a classification task; it represents the prediction whether an annotation is positive or negative with respect to the GO/MeSH sense. We do not assign one of the numerous different senses to a term, but instead a positive or negative label to it, when it corresponds to the GO/MeSH sense or not, respectively. We do not handle acronym ambiguity separately. However, in cases where an acronym belongs to an ontology term label (e.g., FA for GO term 'Fanconi Anaemia' vs 'Fatty Acids', AMP as of MeSH term 'Adenosine Monophosphate' vs 'Antimicrobial Peptides', etc.), this is disambiguated in the same way as all ontology term labels.

As mentioned in the introduction, some disambiguation tasks are easier than others; 'bank' the building and the 'BANK' gene will appear in completely different context, whereas the 'BANK' gene, protein or mRNA are even likely to appear in the same article abstract, making the disambiguation

[8] See http://sourceforge.net/projects/maxent/ for Maximum Entropy implementation

	Term	Senses
GO	Development	biological process of maturation (GO); development of a syndrome/disease/ treatment; cataract development; colony development; development of a method; staff/economic development; software/algorithm development
	Spindle	mitotic spindle (GO); sleep spindles; muscle spindle; spindle-shaped cells
	Nucleus	cell nucleus (GO); body structure (UMLS, subthalamic/cochlear/caudate nucleus); aromatic nucleus
	Transport	directed movement of substances into/out of/within/between cells (GO); patient transport (UMLS); transport by air;transport of virus cultures; maternal transport
MeSH	Thrush	Oral Candidiasis (MeSH); songbird (e.g., thrush nightingale)
	Lead	heavy metal (MeSH); lead measurement (UMLS); to result in
	Inhibition	psychological/behavioral inhibition (MeSH); metabolic inhibition (UMLS); % inhibition (SNOMED)

Tab. 3.11: Ambiguous terms and their senses in the WSD datasets collected. Examples of the senses (in and out of the taxonomies) per ambiguous term included in the benchmark dataset collected.

task often difficult even for a domain expert. 'Transport by air' or 'patient transport' will be easier to distinguish from the GO sense of transport, but 'transport of virus cultures' will appear in a closer molecular biology context. Distinguishing between 'transport', 'RNA transport', 'tRNA transport' or 'ion transport' can become less difficult by using the hierarchical information in the ontology (e.g., exploiting subClassOf/subPropertyOf relations between ontology terms). Some terms are also easier to disambiguate in the same task, depending on the number of their different senses (see Table 3.11) and the distance between them, the way they appear in text (e.g., some can be easily distinguished with the help of regular expressions) and the number of tokens they consist of (one-token terms are usually more ambiguous as they are more likely to correspond to common English).

The ambiguous terms examined are the GO terms 'Development' (GO:0007275), 'Spindle' (GO:0005819), 'Nucleus' (GO:0005634) and 'Transport' (GO:0006810) and the MeSH terms 'Thrush' (D002180), 'Lead' (D007854) and 'Inhibition' (D007266). Most of the different senses of the terms examined (see Table 3.11) belonged as well to the biomedical domain, making the disambiguation task more difficult (e.g., development of a cell culture, development of a cytopathic effect, maturation–GO development). The limited number of terms examined is due to the labor-intensive process of manual collection of proper benchmark datasets. As mentioned in the introduction, in contrast with corpora for the general problem of disambiguation, there exist few annotated biomedical corpora for evaluation and depending on the task, researchers need to collect their own gold standard datasets. We[9] collected datasets for a list of ambiguous terms based on the amount of true/false data available and the frequency of occurrence in PubMed (2600 manually curated documents of high/medium curation quality for 7 selected GO and MeSH terms). We aimed at keeping the ratio of true/false abstracts close to 1, giving a 50% chance to each appearance of the term to be true or false with respect to the GO/MeSH sense (although the ratios in Medline will be different per term). We first examined the UMLS WSD collection (Weeber *et al.*, 2001) for ambiguous GO/MeSH terms and data availability and later a list of common False Positive terms based on manual curations in GoPubMed (terms that were often falsely annotated by GoPubMed as GO/MeSH terms and curators disagreed with the automatic annotation). From the UMLS WSD collection we selected terms that were GO/MeSH terms and the senses provided were distant to each other, i.e. in the case of 'lead', the two senses with short semantic distance (compound; laboratory procedure of lead measurement) were considered as one, as they both are about the compound. A semantically more distant sense is that of the verb to lead/result in. Regarding the false/positive ratio limitation/criterion,

[9]The collection of the datasets has been completely done by the author of this PhD thesis. For purpose of smoother reading, the person is not changed from "we" to "I". Whenever people assisted in the collection process, this is mentioned accordingly.

for some terms this was not satisfied, not allowing the inclusion into the evaluation dataset. For example, for 'transport' the UMLS WSD collection contained 93 abstracts classified as $sense_1$ (True for GO sense) and only 7 as other (curators in this collection had 3 options: $sense_1$, $sense_2$ or other, here $sense_1$ as the biological transport and $sense_2$ as patient transport). We therefore needed to manually collect False examples containing other senses for a balanced corpus.

Another question was whether the definition of the negative datasets would influence the results. To test this, we defined a more general negative dataset by completely randomly choosing articles. Defining a random set as a negative set is common practice, e.g., in predicting protein-protein interactions. Obviously, the random negative dataset is very different from the positive dataset, since most likely it does not contain any of the negative senses at all, but is just the bias for the "average" paper. Results showed a decrease of ∼7% in the performance of the methods.

This argument can be turned around. While our initial negative dataset was carefully and manually chosen, it could be further improved by letting its composition of other senses reflect the distribution of use of these senses in PubMed as a whole. However, achieving this ideal would require annotating all articles with the term in the whole of PubMed with the senses. Given that PubMed has for example more than 1 million articles on development, this cannot be easily accomplished.

However, the composition of negative senses in our negative dataset aims to reflect the composition of negative senses in PubMed as well as possible through the query and annotation strategy, that was pursued. Since we needed to include every possible sense of the ambiguous terms, the queries formed were such that could collect representative abstracts for each sense, a process that was manual and time-consuming. The collection of the positive examples was easier, since there was one sense (with respect to the taxonomy) and also more frequent in PubMed, therefore the term itself or one of its synonyms were enough to be put in the query to PubMed. The collection of negative examples was as expected harder, since they were not frequent in PubMed and we needed to include enough examples of every possible sense. Most of the queries used for this included the ambiguous term or synonyms of it and keywords that were often in the context, based on personal experience from previous curation of automatic annotations in GoPubMed. For example, for 'development', the queries used were 'development AND staff', 'development AND algorithm', 'development AND software', 'development AND treatment', 'development AND method', etc. For 'thrush', since we could only locate one negative sense, we used queries such as 'thrush nightingale', 'thrush AND songbird', 'mountain thrush', etc.

The other aspect is the question of size composition of positive and negative. We chose roughly 50% positive and 50% negative. This basically means that the a priori likelihood is 50% for the corect sense. If, instead, we aim to identify each sense correctly, the following problem arises: assume there are ten senses, i.e. 1 positive and 9 negative. Then the a priori probability for the classification would be 10% and then a simple strategy would be to always vote negative.

Overall, the approach pursued (manual selection of negative senses, roughly covering the common negative senses) plus equally weighted positive and negative datasets is a suitable approach for evaluation.

Datasets

We collected three different benchmark datasets (see Table 3.12) to evaluate the performance of the three methods. They differ in quality and quantity, depending on their collection process (manual by one curator, directed manual by several curators, mainly automatic). The common reference dataset between the three methods is the *manually annotated by a domain expert* one:

High quality, low quantity corpus: this corpus consists of ∼100 true and 100 false example documents (abstracts) per ambiguous term. For the ambiguous GO terms examined and the MeSH term thrush we collected both true and false examples *manually*. True examples are abstracts that discuss, for instance, 'Development', in the sense specified by GO. False examples also contained the ambiguous term, but in other senses, closer or not (see Table 3.11). For the ambiguous MeSH terms 'Lead' and 'Inhibition

	Term	Manual (expert)		Manual (non-experts)		Semi-automatic	
		False	True	False	True	False	True
GO	Development	98	111	271	56	2296	715
GO	Spindle	50	48	70	48	519	599
GO	Nucleus	99	100	25	61	131	1336
GO	Transport	102	91	102	56	1043	699
MeSH	Thrush	17	83	45	7	35	1131
MeSH	Lead	71	27	202	22	1564	735
MeSH	Inhibition	98	100	454	79	5247	553

Tab. 3.12: Benchmark datasets for WSD. The above datasets contain *manually* collected PubMed articles by one expert (high quality / low quantity), *manually* curated articles by a group of non-experts (medium quality / medium quantity) and *semi-automatically* collected articles (low quality / high quantity). See Datasets section for details.

(psychology)', the test set originated from the UMLS WSD corpus (Weeber *et al.*, 2001). These two were the only terms depicting MeSH terms. All other terms in the UMLS WSD (such as growth, repair and reduction) were only found in GO or MeSH as substrings and would thus not be contained in either co-occurrence graph as single nodes.

Medium quality, medium quantity corpus: this corpus consists of documents for which the annotation has been *manually confirmed* by a group of expert and non-expert curators. We asked colleagues to confirm or reject the automatic annotations (for GO and MeSH terms) provided by GoPubMed for a collection of article abstracts. This collection has been mainly automatically created, as described next (low quality, high quantity corpus). For each of the automatic annotations, the curators could select among three options: *a*. true and important for the context of the publication, *b*. of minor importance/relevance and *c*. false annotation. The curation tool is available via GoPubMed (Doms and Schroeder, 2005) (see Figure 3.7).

Low quality, high quantity corpus: this corpus was created mainly *automatically*. We implemented similarity-based clustering of abstracts with literal occurrence of the ambiguous terms. Each abstract was matched to its nearest abstract, conceptualized as a directed edge from the former to the latter. Then every connected component was considered as a cluster. From an initial manual evaluation of the clustering results, clusters of size > 60 were consistent enough, meaning that articles in such clusters were referring to one sense of the ambiguous term in 72-95% of the cases. Each cluster's abstracts were input into a system developed in-house (also used in Alexopoulou *et al.* (2008b)) that generated a list of terms describing each cluster based on term frequency inverse document frequency (TFIDF). The top 20 terms of the list were later evaluated by an expert which labelled the clustered articles as true or false for the respective GO/MeSH term. The above facilitated and accelerated the dataset collection process without any significant loss in data quality (compared to the gain of data quantity for benchmarking).

Experiment

For evaluation and comparison purpose, each method's performance was tested (in terms of precision, recall and specificity) on the high quality / low quantity dataset (see CS1-2, TC1-4 and MD1-3 in Table 3.13 and Tables 3.14, 3.15, 3.17, 3.18 and 3.19 for specificity and detailed results per method). We also applied classical stem co–occurrence analysis as a baseline on the same dataset; this consisted of basic maximum entropy modelling on stems without any use of metadata or hierarchical information (see bME in Table 3.13, Table 3.16 for more details).

We additionally tested each method's performance separately with different test datasets. For the 'Term Cooc' method (TC), the performance of co-occurrences of GO/MeSH terms and inferred co-occurrences of GO/MeSH terms (each one of the variants combined -or not- with Support Vector Ma-

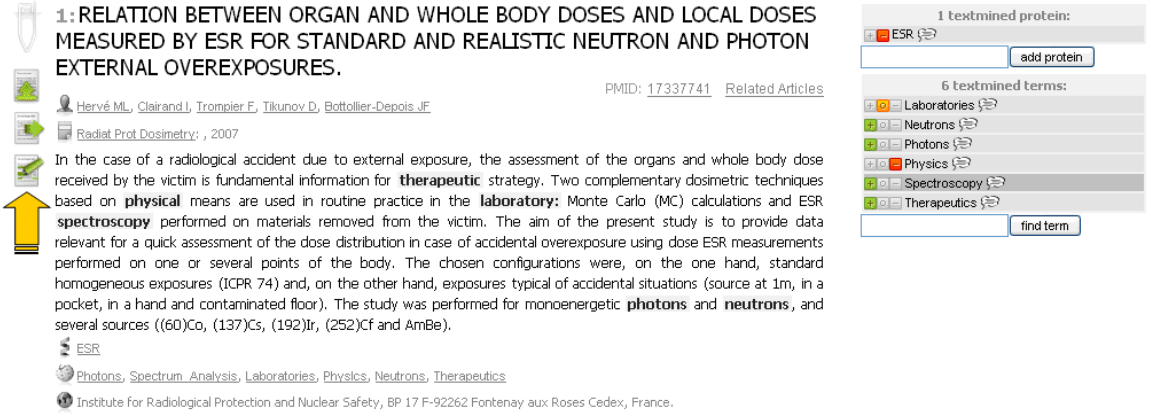

Fig. 3.7: Curation tool in GoPubMed. End-users can confirm or reject the automatic annotations (for GO and MeSH terms) provided by GoPubMed for article abstracts.

chines) was tested in the three benchmark datasets described earlier, in order to evaluate the method in larger (but of lower quality) datasets, since it has been shown that sample size, sense distribution and degree of difficulty impact on the classification task (Xu *et al.*, 2006) (see results in Tables 3.20, 3.21). Input to this method were the automatic annotations per article provided by GoPubMed (GO/MeSH terms and MeSH hand annotation) and the respective co-occurrence graph. As a side experiment, we tested the TC method for the disambiguation of MeSH terms without including the MeSH hand annotations in the automatic annotations provided by GoPubMed, to estimate how the quality of the input influences the quality of the results (see results in Table 3.22). For the 'Closest Sense' method (CS), input was the UMLS semantic network and the article abstracts. This method was additionally tested on the WSD Test Collection (Weeber *et al.*, 2001) (see results in Table 3.23). For the 'MetaData' method (MD), we used the three different datasets for training and testing. The high quality dataset was used in an initial experiment (MD1, detailed results in Table 3.17) as training and testing dataset, in a 5-fold cross validation. Then the medium quality and low quality datasets were separately used as training sets, with testing of the method on the high quality dataset (MD2 and MD3, detailed results in Tables 3.18 and 3.19, respectively).

3.3.3 Results

The performance of the three disambiguation approaches (CS, TC, MD) and the baseline (bME) was tested on a common high quality / low quantity dataset. The overall results of this comparison are shown in Table 3.13 (detailed results per method are given in Tables 3.14, 3.15, 3.16, 3.17, 3.18 and 3.19). All methods perform well between 73-96% average f-measure. In particular, the MetaData (MD1-3) approach is the best one: when trained on high quality data (MD1), it achieves 96% f-measure. When the metadata are not used (baseline method, bME) the accuracy falls to 90%. The Term Cooc (TC1-4) method follows with 81% and the Closest Sense (CS) approach with 77% (80% for the optimized signature together with the subsumption distance, in CS2). All methods present low f-measure for 'development' and 'lead' (79% and 60% in average). The best results (in average for all methods) are obtained for GO terms 'transport', 'nucleus' and 'spindle' (88%, 87% and 85% respectively).

As far as the Closest Sense approach is concerned, there is a clear improvement in the results (from CS1 to CS2) with the use of the optimized signature together with the subsumption distance. For the Term Cooc approach, when the inferred co-occurrences are taken into account (the scores are propagated to the parents of the terms, from TC1 to TC2) in the case of the GO terms the results remain the same, whereas in the case of the MeSH terms the results are worse, mostly in terms of recall (see Table 3.15). For GO terms, the results are best when inferred co-occurrences are combined with SVMs (TC4, 79-98%

Term	CS1	CS2	TC1	TC2	TC3	TC4	bME	MD1	MD2	MD3	Avg
Development	87	86	74	71	57	79	90	96	80	80	80
Spindle	70	79	90	80	95	98	98	100	77	78	87
Nucleus	89	94	81	78	75	95	97	99	91	77	88
Transport	83	71	90	89	88	94	89	98	91	88	88
Thrush	88	94	87	82	78	81	82	94	94	58	84
Lead	36	53	89	49	93	81	85	85	36	14	62
Inhibition	66	84	77	62	85	58	92	100	95	97	82
Avg	74	80	84	73	82	84	90	96	81	70	81

Tab. 3.13: Results (% f–measure) for the baseline (bME) and the three methods (Closest Sense, Term Cooc, MetaData) for 7 ambiguous terms, tested on a high quality / low quantity corpus (manually annotated by expert). CS1 column contains the results (% f-measure) for the Closest Sense (CS) approach with the use of the classic distance (only subsumption). CS2 column contains the results for the CS approach with the use of the optimized signature together with the subsumption distance. TC1 and TC2 contain the results of the Term Cooc (TC) approach, when the co-occurrences or the inferred co-occurrences are used, respectively. TC3 contains the results for the TC approach with co-occurrences and support vector machines, and TC4 when inferred co-occurrences and SVMs are used. bME column contains the results for the baseline method (classical Maximum Entropy modelling of stems without metadata or hierarchical information), trained and tested on the high quality corpus in a 5-fold cross validation. MD1 is for the MetaData approach, trained and tested on the high quality corpus in a 5-fold cross validation. MD2 is trained on the medium quality/quantity corpus and tested on the high quality one. MD3 was trained on the low quality / high quantity corpus and tested on the high quality corpus. Some terms (spindle, nucleus, transport) are easier to disambiguate than others (development, lead). Overall, all methods perform well between 73-96% f-measure (f-measure, F, is the weighted harmonic mean of precision, P and recall, R: $F = 2 \times P \times R/(P + R)$).

f-measure), whereas for MeSH terms, the best f-measure (79-93%) is achieved when co-occurrences with SVMs are used, without the inferred co-occurrences (TC3). This difference can be explained by the different structure of the two hierarchies. GO can be described as "tall and thin" (few children per node, many levels, with maximum number of levels 19), whereas MeSH is "short and fat" (many children per node, not many levels, with a maximum of 9 for the version of 2007). Additionally, the relations between terms in MeSH are not exact is_a relations, but rather is_related_to. Therefore, propagating the term co-occurrences in MeSH does not improve the results, since it does not necessarily mean that annotating with term $MeSH_X$ also means all of X's ancestors. On the contrary, in GO this is more likely to hold. The MetaData method gives - as expected - the best results. When the method is trained and tested on the same high quality test (with a 5-fold cross validation, see MD1), it results in an average f-measure of 96%. When trained on the medium quality (MD2) and low quality (MD3) corpora and tested against the high quality corpus, the f-measure decreases into 81% and 70%, respectively, which are nonetheless high, compared to the quality of the training sets. The high performance of the MetaData approach is mainly due to the use of metadata as the title of the abstract and the journal. For example, for the terms 'inhibition' and 'spindle' it achieves 100% f-measure and for 'nucleus' 99%. The true sense of inhibition for MeSH is psychological inhibition, which is easier to disambiguate, since it will mostly appear in psychology/psychiatry journals. The same applies for 'spindle', which will mostly occur in cell biology and cell division/cycle journals.

We additionally tested each method's performance separately with different test datasets. The 'Closest Sense' method was also tested on the NLM UMLS WSD Collection (Weeber *et al.*, 2001) to compare four versions of semantic distance computation in order to disambiguate term mapping to the UMLS semantic network (see results in Table 3.23). The experiment showed that the use of the ontology definition can improve significantly the precision. Over the 22 ambiguous terms examined, the overall average precision was 83%.

For the 'Term Cooc' method, the performance of the different variants (co-occurrences +/- inferred co-occurrences +/- SVMs) was tested in the three benchmark datasets described earlier (see Datasets section and results in Tables 3.15, 3.20 and 3.21), in order to evaluate the method in larger (but of

lower quality) datasets. Testing the method from the highest towards the lowest quality (but higher quantity), the f-measure decreases only by 3-10%, indicating a consistent behavior of the method. As a side experiment, we tested the 'Term Cooc' method for the disambiguation of MeSH terms without including the MeSH hand annotations in the automatic annotations provided by GoPubMed, to estimate how the quality of the input influences the quality of the results. As expected, the results decreased dramatically (∼46%), indicating that the MeSH hand annotations provided per article are important for the disambiguation (see Table 3.22).

3.3.4 Discussion

Overall, the MetaData method gave the highest f-measures among all methods. The results became worse for the medium and high quantity datasets, since these were of lower quality in terms of correctness. The MetaData approach's consistency with giving the highest results is due to integrating metadata, such as journal and title, which are representative of the true meaning of an ambiguous term. The MetaData approach needs plenty of labelled data for training.

When comparing the results of the Term Cooc and Closest Sense methods to the baseline method (bME) that performs only maximum entropy modelling of stems (without use of metadata), bME still gives better results, but this is due to the available training data of high quality. The disadvantage of MetaData and bME compared to Term Cooc and Closest Sense is the need of high quality training data.

The MetaData approach is less scalable in terms of storage demands as the number of articles increases, while the Closest Sense and Term Cooc approaches have constant storage demands (ontology and a co-occurrence graph).

In the Term Cooc method the SVMs increase the results up to 98%. The Term Cooc method requires an ontology and co-occurrence graphs. The origin of this graph should be a manually curated data source, in our case GOA and MeSH. The quality of the graph will heavily depend on its origin and quality of the data.

The inferred co-occurrences improve the results for GO, while for MeSH they get worse. This is due to the different structures of the two semantic hierarchies; the ancestors of an applicable GO term are more likely to also be applicable to the same article, because of GO's structure that is "tall and thin". But MeSH's structure is "short and fat" and is not always a thesaurus; not all of a node's ancestors are also applicable.

Moreover, in the Term Cooc method the inferred co-occurrences only improve the result if combined with the SVM. This is because the inferred co-occurrences make the extreme co-occurrences with the ambiguous term, which the SVM uses for training, more representative of an ambiguous term's true meaning. Figure 3.8 shows that the most extreme co-occurrences with the ambiguous term are most likely to be classified correctly, since the inferred co-occurrences make more precise the highest and lowest co-occurrences with an ambiguous term. The middle co-occurrences are not necessarily made more precise with inferred co-occurrences. That is why inferred co-occurrences help with the (initial) SVM training; while later on for middle co-occurrences the errors accumulate.

The Closest Sense approach needs only a semantic hierarchy in the form of an ontology, and in this sense is the most automated of the three methods. Moreover, Closest Sense gives good results, where the only problematic term is 'lead'. However, Closest Sense is sensitive to the design of the ontology or subdomain of UMLS used, which reflects the view of the designers. As shown by the accuracy of Humphrey *et al.* (2006) and Liu *et al.* (2004), UMLS may not be the best choice to be used as background knowledge as the different parts of the hierarchy are modelled differently (MeSH, GO, SNOMED, etc.), resulting in different granularity. Different groups of people design ontologies differently; the various subdomains of an ontology will reflect the designers' views respecting depth, number of nodes, and structure. Therefore, the subdomains of the ontology influence the performance of the Closest Sense method, and the design rationale of the ontology may be ultimately responsible

Term	Classic distance (only subsumption)								Optimized signature + subsumption distance							
	Threshold 0.8				Threshold 0.7				Threshold 0.8				Threshold 0.7			
	P	R	S	F	P	R	S	F	P	R	S	F	P	R	S	F
Development	0.40	0.94	0.59	0.56	0.98	0.78	0.97	0.87	0.98	0.55	0.83	0.71	0.77	0.98	0.79	0.86
Spindle	0.65	1	0.75	0.78	0.90	0.57	0.78	0.70	0.98	1	0.98	0.99	0.96	0.67	0.93	0.79
Nucleus	0.63	0.96	0.56	0.76	0.97	0.82	0.90	0.89	0.94	1	0.88	0.97	0.98	0.91	0.94	0.94
Transport	0.24	0.88	0.46	0.38	0.87	0.80	0.78	0.83	0.32	0.97	0.50	0.48	0.81	0.64	0.54	0.71
Thrush	0.62	1	0.38	0.76	0.89	0.88	0.50	0.88	0.97	0.99	0.89	0.98	1	0.89	1	0.94
Lead	0.30	0.57	0.77	0.39	0.41	0.32	0.75	0.36	0.30	0.57	0.77	0.39	0.74	0.42	0.86	0.53
Inhibition	0.52	0.70	0.62	0.60	0.89	0.52	0.60	0.66	0.85	0.96	0.86	0.90	1	0.73	1	0.84

Tab. 3.14: High quality / low quantity corpus: Precision/ Recall/ Specificity / F–measure for the Closest Friends (CF) method on the GO and MeSH test datasets.

Term	Co-occurrences				Hierarchical Cooc				Cooc + SVMs				Hierarchical Cooc + SVMs			
	P	R	S	F	P	R	S	F	P	R	S	F	P	R	S	F
Development	0.74	0.74	0.31	0.74	0.70	0.72	0.20	0.71	0.61	0.54	0.91	0.57	0.79	0.78	0.85	0.79
Spindle	0.91	0.90	0.79	0.90	0.83	0.78	0.49	0.80	0.95	0.94	1	0.95	0.98	0.98	1	0.98
Nucleus	0.81	0.80	0.10	0.81	0.80	0.75	0.07	0.78	0.77	0.73	0.90	0.75	0.95	0.95	0.90	0.95
Transport	0.90	0.90	0.60	0.90	0.89	0.89	0.30	0.89	0.90	0.87	0.99	0.88	0.94	0.93	1	0.94
Thrush	0.92	0.83	0.94	0.87	0.82	0.82	0.72	0.82	0.89	0.70	1	0.78	0.89	0.74	1	0.81
Lead	0.89	0.89	0.94	0.89	0.81	0.35	0.50	0.49	0.93	0.92	0.93	0.93	0.80	0.81	0.95	0.81
Inhibition	0.78	0.76	0.62	0.77	0.72	0.55	0.35	0.62	0.85	0.85	0.65	0.85	0.59	0.56	0.85	0.58

Tab. 3.15: High quality / low quantity corpus: Precision / Recall / Specificity / F–measure for the Term Cooc (TC) method on the GO and MeSH test datasets.

Term	neg	pos	P	R	S	F
Development	111	98	0.84	0.98	0.81	0.9
Spindle	50	48	1	0.96	0.96	0.98
Nucleus	99	200	0.98	0.96	0.96	0.97
Transport	102	91	1	0.81	0.75	0.89
Thrush	17	80	0.77	0.88	0.87	0.82
Lead	71	27	0.96	0.77	0.71	0.85
Inhibition	98	100	0.93	0.92	0.91	0.92

Tab. 3.16: High quality / low quantity corpus: Precision / Recall / Specificity / F–measure for the baseline (bME) method on the GO and MeSH test datasets. 5-fold cross validation on the high quality/low quantity corpus.

Term	neg	pos	P	R	S	F
Development	98	111	0.92	1	0.91	0.96
Spindle	50	48	1	1	1	1
Nucleus	99	200	1	0.99	0.99	0.99
Transport	102	91	0.96	1	0.96	0.98
Thrush	17	80	1	0.88	1	0.94
Lead	71	27	0.74	1	0.64	0.85
Inhibition	98	100	1	1	1	1

Tab. 3.17: High quality / low quantity corpus: Precision / Recall / Specificity / F–measure for the MetaData (MD) method on the GO and MeSH test datasets. 5-fold cross validation on the high quality/low quantity corpus.

Term	neg train	pos train	neg test	pos test	P	R	S	F
Development	271	56	98	111	0.67	0.99	0.52	0.80
Spindle	70	48	50	48	0.63	1	0.42	0.77
Nucleus	25	61	99	200	0.85	0.98	0.83	0.91
Transport	102	56	102	91	0.84	1	0.80	0.91
Thrush	42	5	17	80	1	0.88	1	0.94
Lead	202	22	71	27	1	0.22	1	0.36
Inhibition	454	79	98	100	0.96	0.94	0.96	0.95

Tab. 3.18: MetaData (MD) method results: Training on medium quality/medium quantity corpus, testing on high quality/low quantity corpus.

Term	neg train	pos train	neg test	pos test	P	R	S	F
Development	2296	715	98	111	0.67	1	0.50	0.80
Spindle	519	599	50	48	0.65	1	0.46	0.79
Nucleus	131	1336	99	200	0.63	1	0.40	0.77
Transport	1043	699	102	91	0.80	1	0.74	0.88
Thrush	35	1131	17	80	1	0.41	1	0.58
Lead	1564	735	71	27	1	0.07	1	0.14
Inhibition	5247	553	98	100	0.99	0.96	0.99	0.97

Tab. 3.19: MetaData (MD) method results: Training on low quality/high quantity corpus, testing on high quality/low quantity corpus.

Term	Co-occurrences				Hierarchical Cooc				Cooc + SVMs				Hierarchical Cooc + SVMs			
	P	R	S	F	P	R	S	F	P	R	S	F	P	R	S	F
Development	0.76	0.78	0.74	0.77	0.76	0.64	0.66	0.69	0.76	0.77	0.85	0.77	0.78	0.80	0.85	0.79
Spindle	0.76	0.75	0.70	0.75	0.80	0.75	0.62	0.77	0.81	0.77	0.92	0.79	0.81	0.79	0.96	0.80
Nucleus	0.86	0.85	0.51	0.85	0.84	0.84	0.48	0.84	0.80	0.36	0.82	0.50	0.83	0.60	0.93	0.70
Transport	0.89	0.89	0.90	0.89	0.87	0.87	0.87	0.87	0.80	0.75	0.98	0.78	0.83	0.78	0.98	0.80
Thrush	0.73	0.71	0.82	0.72	0.77	0.65	0.71	0.71	0.88	0.15	0.11	0.26	0.75	0.83	0.87	0.78
Lead	0.83	0.80	0.87	0.82	0.84	0.76	0.81	0.80	0.82	0.82	0.80	0.82	0.82	0.82	0.80	0.82
Inhibition	0.74	0.80	0.80	0.77	0.74	0.75	0.74	0.75	0.70	0.61	0.61	0.65	0.69	0.54	0.58	0.61

Tab. 3.20: Medium quality / medium quantity corpus: Precision / Recall / Specificity / F–measure for the Term Cooc (TC) method on the GO and MeSH test datasets.

Term	Co-occurrences				Hierarchical Cooc				Cooc + SVMs				Hierarchical Cooc + SVMs			
	P	R	S	F	P	R	S	F	P	R	S	F	P	R	S	F
Development	0.68	0.66	0.55	0.67	0.72	0.60	0.46	0.65	0.70	0.71	0.73	0.70	0.74	0.75	0.83	0.74
Spindle	0.79	0.80	0.63	0.79	0.79	0.78	0.45	0.78	0.83	0.76	0.87	0.80	0.88	0.87	0.87	0.87
Nucleus	0.94	0.81	0.77	0.87	0.96	0.95	0.74	0.96	0.94	0.80	0.99	0.86	0.95	0.90	0.99	0.93
Transport	0.69	0.68	0.61	0.69	0.71	0.72	0.52	0.72	0.71	0.63	0.78	0.67	0.72	0.65	0.86	0.68
Thrush	0.54	0.49	0.89	0.52	0.57	0.55	0.74	0.56	0.91	0.91	0.91	0.91	0.43	0.45	0.80	0.44
Lead	0.55	0.53	0.86	0.54	0.53	0.53	0.80	0.53	0.55	0.53	0.78	0.54	0.49	0.51	0.70	0.50
Inhibition	0.75	0.50	0.85	0.60	0.64	0.51	0.75	0.57	0.34	0.34	0.51	0.34	0.26	0.28	0.37	0.27

Tab. 3.21: Low quality / high quantity corpus: Precision / Recall / Specificity / F–measure for the Term Cooc (TC) method on the GO and MeSH test datasets.

Term	Co-occurrences			Hierarchical Cooc			Cooc + SVMs			Hierarchical Cooc + SVMs		
	P	R	F	P	R	F	P	R	F	P	R	F
Thrush	0.54	0.25	0.34	0.69	0.46	0.55	0.86	0.83	0.85	0.86	0.83	0.85
Lead	0.72	0.75	0.73	0.69	0.73	0.72	0.80	0.35	0.48	0.81	0.35	0.48
Inhibition	0.24	0.47	0.32	0.37	0.48	0.42	0.80	0.67	0.73	0.80	0.67	0.73

Tab. 3.22: High quality / low quantity dataset with MeSH Text-mined annotations only: Precision / Recall / F–measure for the Term Cooc (TC) method on the GO and MeSH test datasets.

Term	Only subsumption	Optimized
adjustment	0.66	0.66
ganglion	0.92	0.93
extraction	0.62	0.65
japanese	0.65	0.92
pressure	0.98	0.86
surgery	0.61	0.6
depression	0.86	0.96
lead	0.64	0.6
radiation	0.94	0.92
sensitivity	0.98	0.74
transient	0.7	0.97
fat	0.64	0.73
growth	0.85	0.85
man	0.87	0.87
sex	1	0.94
cold	0.8	0.95
fit	0.97	0.69
immunosuppression	0.68	0.84
repair	0.65	0.74
condition	0.79	0.8
implantation	0.97	0.99
strains	0.92	0.94
Average precision	0.8	0.83

Tab. 3.23: Results (precision) of the Closest Sense (CS) method tested on the WSD Test Collection (Weeber et al., 2001), with the use of classic distance (only subsumption) and with the use of the optimized signature together with the subsumption distance.

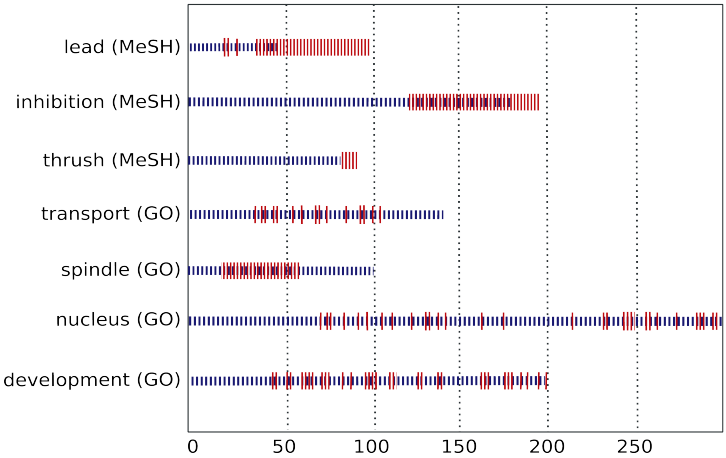

Fig. 3.8: Term Cooc classification over time. The x-axis is the TC classification over time. Left-most articles are classified early, since they have the highest or lowest co-occurrences with the ambiguous term. Red lines represent errors or wrong predictions. Almost none of the early classified articles are errors.

for performance differences on various terms. For example, 'nucleus' is a subtree root in both GO and SNOMED (anatomical structures); in GO there are 2000 descendants of nucleus, while in SNOMED 10.

3.3.5 Conclusion and Future work

Based on the results, metadata and training data of high quality seem to be the key point for the increase of the accuracy. When such training data are not available - as happens in most of the cases - co–occurrence of ontology/taxonomy/thesaurus terms can provide the way to the right decision. Moreover, the hierarchy of the terms and the subdomain, when consistently modelled, can depict the correct sense of an ambiguous term.

The MetaData method produced the best results by including metadata in the WSD decision, but it requires high quality training data. The most interesting thing about the Term Cooc and Closest Sense methods is that they are semi-automated, given a co-occurrence graph or ontology; then the training does not require manual intervention. Term Cooc requires well modelled ontologies such as GO, and deteriorates as the structure becomes less rigorous as in MeSH. Closest Sense requires large and consistently modelled ontologies, which are two opposing requirements. Thus, for TC and CS the structure of the ontology and subdomain affect the distance metric used and WSD quality. Future work could include identifying ambiguous terms for a certain corpus automatically. For this purpose, we could employ WordNet, clustering, Part of Speech and noun phrase statistics, and expert input.

For Term Cooc and Closest Sense, we assumed that the other terms in the context are correct and independent of one another; in fact, they could also be ambiguous and therefore false. For Closest Sense we could optimize the distance computation and propose other distances, taking into account existing annotation bases and ontology structure.

So far, the disambiguation performed was between the true sense in the hierarchy and all other senses that were considered as false. A possible extension of the methods would be to correctly identify if a sense occurs that is not included in the thesaurus/ontology and possibly add it. The Closest Sense method can potentially do this by setting a threshold. From all distances below this certain threshold, one should be clearly shortest. If not, then this indicates a new sense. The Term Cooc and MetaData approaches could be adjusted to identifying new senses by training each method on each sense and setting a certain threshold. If the sense found is not above the threshold, then this can be a new sense.

Another idea would be to combine the three disambiguation approaches (MetaData, Term Cooc, Closest Sense) and get a confidence score for each of the approaches each time a sense is being disambiguated.

So far, we have only worked on abstracts of PubMed articles. It would be interesting to see how the accuracy would change once the disambiguation would be performed in the full text of articles and also the co–occurrences would be computed based on full-text articles instead of abstracts. The context in full text would definitely change from paragraph to paragraph and appearance in the same text of e.g., 'development' as 'heart development' or 'development of a syndrome' is highly likely to happen. In some rare cases, this happens also in the same paragraph or even in the article abstract, as shown for example in Figure 3.9. In such cases, use of Natural Language Processing would be more suitable to resolve the ambiguity, taking into account the exact sentence and the part of speech the ambiguous term belongs to.

We have also tried to create a "negative co–occurrence graph" to be used in disambiguation. This would include the 'worst enemies' (instead of the 'best friends' used so far) of the true sense of the ambiguous term. Once a 'negative cooc' would be found, this would give a negative score additionally to the Term Cooc score and the final score would give the decision for true or false. As already mentioned in Section 3.3.2, corpus collection is tedious and time-consuming, especially that of the negative examples. The corpora created for the evaluation (described in Section 3.3.2) were complete enough to serve as benchmark datasets, but could not serve as a basis for the computation of co–occurrences representative

Title: Bioinformatics analysis of microarray data.
Authors: Zhang Y, Szustakowski J, Schinke M.
Journal: Methods Mol Biol. 2009;573:259-84.

Abstract: **Gene expression profiling provides unprecedented opportunities to study patterns of gene expression regulation, for example, in diseases or *developmental* processes.** Bioinformatics analysis plays an important part of processing the information embedded in large-scale expression profiling studies and for laying the foundation for biological interpretation. Over the past years, numerous tools have emerged for microarray data analysis. **One of the most popular platforms is Bioconductor, an open source and open *development* software project for the analysis and comprehension of genomic data, based on the R programming language. In this chapter, we use Bioconductor analysis packages on a heart *development* dataset to demonstrate the workflow of microarray data analysis from annotation, normalization, expression index calculation, and diagnostic plots to pathway analysis, leading to a meaningful visualization and interpretation of the data.**

PMID: 19763933 [PubMed - indexed for MEDLINE]

Fig. 3.9: Two senses for 'development' in the same article abstract in PubMed: 'Software development' and 'biological development' (and its descendant, 'heart development').

of the whole of PubMed.

Chapter 4

Terminologies for Text-Mining

At present the field of biology faces the problem of the presence of a large amount of data without any associated semantics. Therefore, biologists currently waste a lot of time and effort in searching for all of the available information about each area of research. This is hampered further by the wide variations in terminology that may be in common usage at any given time, and that inhibit effective searching by computers as well as people. In recent years, to facilitate biomedical research, various ontologies and knowledge bases have been developed.

As already mentioned in Section 2.2.3, the engineering of ontologies, especially with a view to a text-mining use, is still a new research field. A well-defined theory and technology for ontology construction does not yet exist. Many of the ontology design steps remain manual and are based on personal experience and intuition. However, there exist several efforts on automatic construction of ontologies in the form of extracted lists of terms and relations between them (Frantzi *et al.*, 2000; Navigli and Verlardi, 2004; Cimiano and Völker, 2005; Zavitsanos *et al.*, 2007; Wächter, 2010).

In this Chapter, we share the experience acquired during the manual development of a lipoprotein metabolism ontology (LMO) to be used for text-mining. We provide guidelines for the design of this ontology and describe the common obstacles during the process. We compare the manually created ontology terms with the automatically derived terminology from four different automatic term recognition (ATR) methods. The top 50 predicted terms contain up to 89% relevant terms. For the top 1000 terms the best method still generates 51% relevant terms. In a corpus of 3066 documents 53% of LMO terms are contained and 38% can be generated with one of the methods. We conclude with a discussion on the automation of the ontology generation process and how this can be best achieved.

The current work on ontology design has been published in Alexopoulou *et al.* (2008b) and has been part of the EU–funded project Sealife[1] (Schroeder *et al.*, 2006).

[1] http://www.biotec.tu-dresden.de/sealife/

4.1 Methods

4.1.1 Ontology Design Principles

The Open Biomedical Ontologies (OBO) Foundry[2] provides ontology design principles concerning the syntax, unique identifiers, content and documentation of ontologies to be added or edited, as a common agreement between users/editors. OBO principles that are not discussed later (but were followed during the Lipoprotein Metabolism Ontology design) refer mainly to the use of a common shared syntax (OBO syntax and extensions or OWL), the insertion of a unique identifier per term, the relations included in the OBO Relation Ontology and the clearly delineated content (terms in different ontologies should provide distinguishable descriptions of a concept). The success of the OBO representation format is attributed to its informal expressivity, combined with the ability of conversion into OWL and vice-versa.

The only OBO principles we did not follow were the free availability and collaboration with other OBO Foundry members, due to corporate reasons. However, we present the knowledge acquired and the problems faced during the ontology design. The following guidelines, as well as the decisions, compromises and problems described later derive from our experience during the manual development of the Lipoprotein Metabolism Ontology.

Some basic steps that should be followed during the design of any ontology include *identifying the range of intended users*, *deciding on the purpose and main research area* of the ontology and defining/predicting further *possible applications* (e.g., GO has also been used by the search engine GoPubMed[3] (Doms and Schroeder, 2005) and by GoMiner[4] (Zeeberg *et al.*, 2003) for gene expression data evaluation, although its initial purpose did not include use for text-mining). Important points to start from are *literature scanning* for deciding on the basic concepts as well as the *insertion of a textual definition for each term*. *Formulation of questions* is also crucial (Uschold, 1996). Examples of questions that researchers from Unilever needed to answer were:

- "What is the activity of cholesterol ester transfer protein (CETP) in diabetes?"
- "Which tissues is apoE expressed in?"
- "What is the impact of fish oil diet in metabolic syndrome patients?"
- "Which tissues is LPL expressed in? How does this expression change in diabetes?"
- "What is the activity of HL in obese individuals?"
- "What is the Km for LPL catalyzed VLDL catabolism?"
- "How does HL work? Does it hydrolyze core lipids or surface only?"
- "What is the preferred substrate of EL?"
- "What is the mechanism of action of CETP? Does it have two lipid binding pockets or one?"
- "How does the lipid composition of LDL differ in diabetes and obese men?"
- "What is the role of PLTP in HDL metabolism?"
- "What is the role of apoE in apoB100-lipoprotein kinetics?",

etc, indicating that terms such as 'CETP', 'diabetes', 'apoE', 'diet', 'fish oil diet', 'metabolic syndrome' and 'patient' should be included in the ontology. *Reusing existing ontologies* that may cover to some extent the ontology under design or could be inserted as a separate branch of the ontology is also a

[2]See OBO foundry http://www.obofoundry.org/
[3]See GoPubMed http://www.gopubmed.org/
[4]See GoMiner http://discover.nci.nih.gov/gominer/

possibility. In the case of the Lipoprotein Metabolism Ontology (LMO), we needed to include information on diet. For this purpose, we included the Nutrition Ontology from the NCI Cancer Nutrition Ontology Project[5] as a separate part under diet. *Deciding on a label for each concept* is one of the most crucial steps during the structuring of the ontology. This task is difficult for humans as it requires good knowledge of the domain of interest so as to group concepts on the hierarchy in a semantically meaningful way. It is even more difficult for machines to do this automatically. There exists previous work on automatic labeling of document clusters (Popescul and Ungar, 2000) by using the most frequent and most predictive words in clusters of documents, but there is still work to be done in this direction. One must firstly concentrate on the semantics of a term, decide what is really needed to be expressed with that term, and then choose the appropriate name. Last, but not least, and perhaps one of the very first steps of the designing procedure is the selection of a suitable *ontology editor*. We used the Protégé OWL plug-in[6] for building the ontology and CmapTools[7] for visualization (see Figure 4.1). Ontology visualization is crucial when the knowledge engineer and the domain expert are two different persons and need to agree on the different versions of the ontology.

With GO we experienced some limitations for text-mining. For example, it is unlikely that a descriptive label such as 'cell wall (sensu Gram-negative bacteria)' will literally appear in text. A comprehensive overview of such problems is provided by Smith *et al.* (2004). There often exist ontology terms that are unlikely to appear in text, and are rather of a structuring nature. For example, the terms 'hydrolase activity, acting on ester bonds' (GO:0016788) or 'hydrolase activity, acting on carbon-nitrogen (but not peptide) bonds' (GO:0016810) include several different types of information: activity (hydrolase), type of bond affected (ester or carbon-nitrogen) and exception (but not peptide) (see Figure 4.2). These should be 3 different branches of the tree, combined with relations, therefore structuring 'logical formulas'. For example, in the case of the second term (GO:0016810), the exception could be expressed as a certain condition: the protein has a hydrolase activity and is acting on carbon nitrogen bonds, but *not in all* bonds (peptide bonds are excluded).

Aranguren *et al.* (2007) provide a simple and indicative example of the problem: a Person is a Man or a Woman, a Man has Testis, a Woman has no Testis, but what happens in the case of a Eunuch (who is actually a man without Testis)? There is a need for distinguishing between relations that are strict "always" rules and "normally" or "usually" relations that can also allow for exceptions. Biomedical terms are usually connected with "usually" relations between them.

Another example is the definition of mammals: a simple definition[8] can be 'warm-blooded vertebrate animals belonging to the class mammalia, including all that possess hair and suckle their young'. Therefore, one can say that all mammals give birth to and suckle their young. But there exists the exception of the monotremes, which are mammals that lay eggs instead of bearing live young. The definition here would be "mammals are animals that *normally* bear live young and suckle them" and the exception "monotremes are mammals that lay eggs".

Another example is given by Hoehndorf *et al.* (2007) (from the Foundational Model of Anatomy), where *"every instance of a human body has as part an appendix"*, corresponding to an idealized (canonical) "normal" human. However, an individual human body may lack an appendix as a part, demonstrating that canonical ontologies do not always represent default knowledge and should include exceptions. Hoehndorf et al. developed a methodology for representing canonical domain ontologies within the OBO foundry by adding an extension to the semantics for relationships in the biomedical ontologies that allows for treating canonical information as default. Rector explored some of the alternatives in OWL and related languages for dealing with issues such as exceptions (predictable and not) and limited expressivity (Rector, 2004). Rector's analysis is divided in four cases, which can be resolved with OWL, more precise

[5] See http://gforge.nci.nih.gov/projects/nutrition/
[6] For the Protégé OWL plug-in, see http://protege.stanford.edu/overview/protege-owl.html
[7] For Concept Map Tools (CmapTools), see http://cmap.ihmc.us/
[8] See http://www.biology-online.org/dictionary/Mammals for definition of mammals

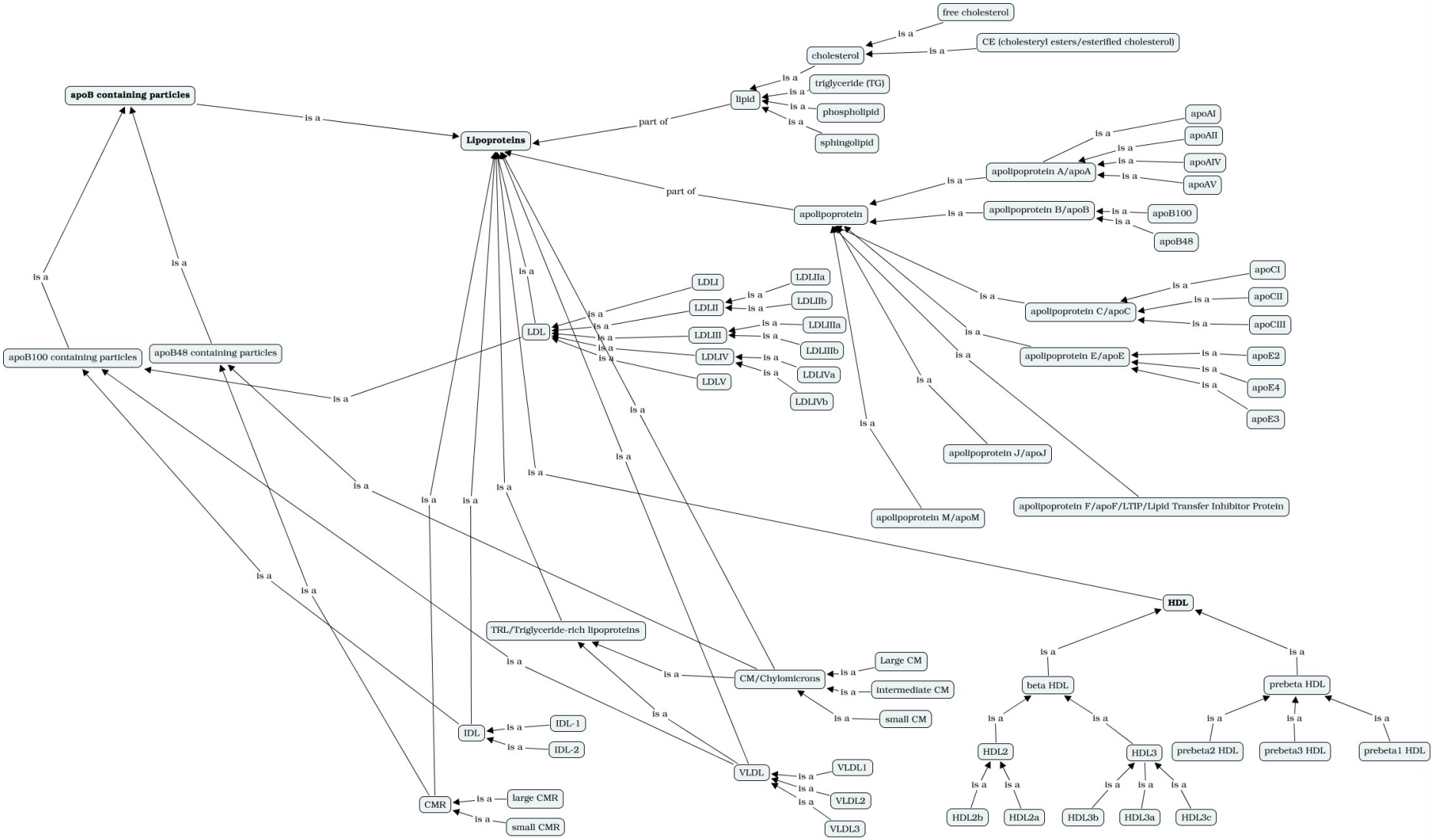

Fig. 4.1: CmapTools representation of a part of the Lipoprotein Metabolism Ontology. This visualization was used from the knowledge engineer and the domain expert in order to agree on the different versions of the ontology.

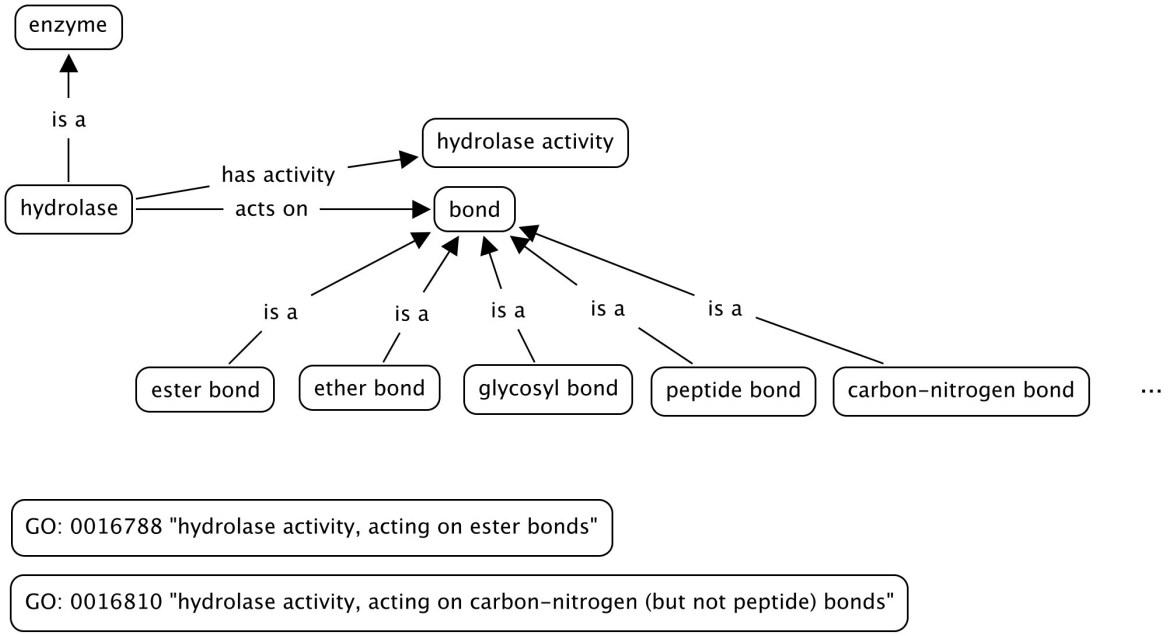

Fig. 4.2: Problematic terms – the hydrolase activity example. Terms like hydrolase, hydrolase activity, bond, ester bond and relations between them (e.g., acts on) can be easily found in text, whereas full GO terms such as 'hydrolase activity, acting on ester bonds' are unlikely to appear literally in an article.

logical formulation (OWL-DL), more explicit context and generalized common information, and other more complicated methods.

Compositional structure of terms is a major bottleneck for ontology design, especially when it comes to text mining, as the relations between terms must be as simple as possible. Ogren *et al.* (2004, 2005) have performed an analysis of the term names in the GO to investigate substring relations between terms and revealed that 65.3% of all GO terms contain another GO term as a proper substring. These terms can be categorized into two groups: GO terms that contain other GO terms as proper substrings (e.g., 'hydrolase activity, acting on acid sulfur-sulfur bonds' (GO: 0016828) and 'hydrolase activity' (GO: 0016787)) and GO terms that contain strings that seem to recur frequently (e.g., 'regulation of' in GO, 'predominance of' in the Lipoprotein Metabolism Ontology). Text-mining ontologies can be extensions of annotation ontologies which enrich annotation ontologies with synonyms suitable for text-mining. Some decisions and compromises have to be made on the relationships and on the labels defined during the concept hierarchy design.

4.1.2 Decisions that need to be made during the ontology design

Keep or dismiss a term: When using the ontology for text-mining over a specific biomedical domain, it is important to include terms specific enough to define the domain and also general enough to cover it entirely. For example, including information on 'kinetics' during the design of the Lipoprotein Metabolism Ontology is crucial. But 'kinetics' is too general as a term, as the distinction between different kinds of kinetics is important (e.g., when querying PubMed for 'kinetics', there are retrieved articles referring to 'kinetics of phenols' or a 'reconstruction kinetics well', irrelevant to the domain of interest). On the other hand, the term 'lipoprotein kinetics' is too specific and documents mentioning it do not cover all essentials known in lipoprotein kinetics. Searches for "lipoprotein kinetics", "lipoprotein" and "kinetics" and retrieval of relevant articles (e.g., PMID: 12606523 'Differential regulation of lipoprotein kinetics by atorvastatin and fenofibrate in subjects with the metabolic syndrome.') lead to the decision that the best term to use for 'lipoprotein kinetics' is the exact term. There already exist previous efforts on automatic labeling of document clusters and identification of ontology components, based on Natural

Language Processing techniques or hierarchical and suffix-tree clustering (Stefanowski and Weiss, 2003; Lame, 2004; Popescul and Ungar, 2000).

Decide on ontology design/relations: the ontology must be a subsumption ontology. It can be either a structured vocabulary/terminology containing only child-parent relationships (mostly 'is_a' and 'part_of') between concepts or an ontology of different complexity that could be easily translated into a simple hierarchy. It should also be rich in synonyms and textual definitions (mentioned earlier), that would be useful for disambiguation. For the Lipoprotein Metabolism Ontology we used the Protégé OWL plug-in, with concepts being the term labels (e.g., human / Tangier's disease) and instances the term synonyms/variants (e.g., patient, test person, experimentee / analphalipoproteinemia).

Decide on synonyms: researchers do not have strict and formal ontologies or nomenclatures in their minds when composing a scientific article and therefore use terminology of differing granularity. They often use parent terms to refer to a child term, or vice-versa (e.g., 'coronary artery disease (CHD, CAD)' is child of 'cardiovascular disease', but in many cases authors are treating them the same). Again literature scanning, for both child and parent term, will help to clarify how researchers refer to different terms. Another problematic case is that of the different lipoprotein subclasses (based on particle size, buoyant density, composition, etc.) where there do not exist clear limits between them. Depending on the way of measurement and the difference in surface lipid content, they can be expressed in different ways. For example, in the case of LDL, there are 5 different subclasses based on particle size (LDL I-V), but there are also references such as 'small dense LDL' or 'buoyant LDL' that are very often found in text but could contain a mixture of different subclasses. Since we need to keep only a simple hierarchy with parent-child relationships, we do not incorporate any "definitional" information (e.g., that 'small dense LDL' consists of a mixture of LDLIII and LDLIV). In these cases, we put the synonyms according to the authors' use; for example 'small dense LDL' as a synonym for LDL III and 'buoyant LDL' or 'large LDL' as synonyms for LDL I (Berneis and Rizzo, 2005). A similar example from the GO is that of 'transporters and carriers'. In every day language 'transporter and carrier' is the same as 'transporters or carriers', but they are logically different.

Handle term variation: terms like 'Tangier disease', 'Tangier's disease' or 'Tangiers disease' are variants of the same term. Terms like 'LDL I', 'LDL-I', 'LDL-1', 'LDL 1', 'LDL1' and 'LDLI' are also variants of the same term. The process of manually inserting such lexical variants (with hyphens, apostrophes, slashes, or even American/British spelling variants) in the ontology is tedious and time-consuming. There exist programs that handle such variations by producing the normalized form of a term, such as the UMLS Lexical Variant Generation Program (McCray *et al.*, 1994). For the Lipoprotein Metabolism Ontology we did not use such a variant generation program. We included only term variants we could find in literature.

4.1.3 Compromises that need to be made, Problems, Inconsistencies

Some compromises must be made to retain a correct ontology (meaning that it contains valid relations) and still get the best possible results from text-mining:

Ambiguity resulting either from identical *abbreviations* for different terms (e.g., 'CAM' can stand for 'constitutively active mutants', 'cell adhesion molecule', or 'complementary alternative medicine'), or *ambiguous term labels* (e.g., 'embryo' for 'mouse embryo' or 'male' for 'male patients') is always a problem. Abbreviations and acronyms should be included in the ontology, but conservatively or with an appropriate algorithm that could handle them. Word sense disambiguation is a salient point here; knowledge sources like long-form/short-form combinations, domain (context under which the word is

used) and collocations (adjacent words/terms) can be exploited to provide the correct sense of the term (Agirre and Stevenson, 2006). For the case of incomplete term labels, let us consider the following example: we are only interested in experiments performed in human patients and need to distinguish between human- and animal- referring articles. One option is to insert into the ontology only human-specific terms, such as 'experimentee', 'patient', 'man', 'boy', etc. 'Male' cannot be in the ontology, since it could also be referring to animals. Another option is to maintain a list of human- and animal- specific words or expressions and then transform the algorithm in a way that one could make a Boolean selection (e.g., AND human, NOT animal) in the query and finally include or exclude the results for the specific selections.

Try to avoid any possible inconsistencies. To illustrate the implication of inconsistencies on reasoning, let us describe the following example: a researcher is interested in the different lipoprotein levels in patients of different race and geographical location, since there has been evidence that these two factors affect lipoprotein metabolism. Combination of geographical information as well as racial information in one part of the ontology is, therefore, needed. Many articles refer to "African-Americans" as "blacks", so the term must be included under 'ethnic group'. Then the following must be valid:

- define 'Caucasian', 'African' and 'Asian' as 'ethnic group'
- 'American' *is a* 'Caucasian'
- 'African-American' *is a* 'African'
- 'African-American' *is a* 'American'
- 'African-American' is 'black' (*synonym*)
- 'Caucasian' is white (*synonym*)
- **but** 'African-American' cannot be 'Caucasian' or 'white' (although he is 'American').

This is similar to the case of mammals that lay eggs or the 'Man, Woman, Eunuch' example described earlier in Section 4.1.1; people very often formulate rules such as *"normally is-a"*, as there are always exceptions. For the LMO we excluded the 'American' concept and added 'African-American' as child of 'African' and 'Hispanic-American' as child of 'Caucasian'.

4.2 Results

The Lipoprotein Metabolism Ontology (LMO) was manually built in collaboration with domain experts from Unilever for the purpose of document retrieval. It consists of 223 concepts and 623 additional synonyms, with an average term length of 14 (2 words of 7 characters). A concept as used here consists of a concept label and optional synonymous terms. A term can be any word or phrase of relevance to the studied domain. Together with the Nutrition Ontology from the NCI Cancer Nutrition Ontology Project[9], the LMO contains in total 522 concepts and 964 additional synonyms, with an average term length of 15 (2 words of 7.5 characters). Concerning the relations between the concepts, the mean number of parents is 2 (with a maximum of 3) and the mean number of siblings is 5 (with a maximum of 10). We did not include the Nutrition Ontology terms in the experiment, as we only wanted to compare the terminology created manually by us with the automatically derived terminologies from the different term extraction methods.

[9]See http://gforge.nci.nih.gov/projects/nutrition/

For Automatic Term Recognition (ATR), a 'lipoprotein metabolism'-specific corpus was created, consisting of 300 abstracts collected from PubMed with the query "lipoprotein metabolism" (limit for Review papers). These 300 abstracts were the maximal number of articles where all methods delivered results. Five different ATR methods were tested on that corpus, namely Text2Onto, OntoLearn, Termine (Cimiano and Völker, 2005; Navigli and Verlardi, 2004; Frantzi *et al.*, 2000) and two methods developed in-house, one considering the relative frequency (RelFreq) of a term in the corpus and the other (TFIDF) additionally using the document frequency derived from all phrases contained in NCBI's PubMed database.

Termine (Frantzi *et al.*, 2000) extracts noun phrases and for ranking it considers several statistical characteristics of the candidate term, such as the total frequency of occurrence in the corpus, the frequency of the term as part of other longer candidate terms (and the number of these) and the length of the candidate term (in number of words).

Text2Onto (Cimiano and Völker, 2005) also extracts noun phrases and is based on algorithms calculating the Relative Term Frequency and TFIDF, as well as Entropy and the C-value/NC-value used by Termine in order to extract the concepts. It further exploits the hypernym structure of WordNet (Fellbaum, 1998), matches Hearst patterns (Hearst, 1992) and others in the corpus in order to get the relations (subclass_of, part_of, instance_of), but at this point we only examined the terminology extraction precision.

OntoLearn (Navigli and Verlardi, 2004) uses a linguistic processor and a syntactic parser in order to extract a list of syntactically plausible terminological noun phrases. For filtering "true" terminology, OntoLearn is based on two measures, namely Domain Relevance and Domain Consensus, which calculate the specificity of a candidate term with respect to the target domain via comparative analysis across different domains as well as the distributed use of a term in a domain. OntoLearn consists of a full ontology learning pipeline, starting from extracting the terminology, building the hierarchy and creating definitions for the terms.

The **in-house method** extracts noun phrases in a way similar to Termine. The ranking performed is domain-specific, using term frequency and global frequency in all 17 million abstracts from PubMed. The first version of the method considers the relative frequency (**RelFreq**) of a term in the corpus and the second version (**TFIDF**) additionally uses the document frequency derived from all phrases contained in PubMed. Terms of syntactic similarity are grouped together and abbreviations are linked to their long forms. For more details, see (Wächter, 2010).

OntoLearn was excluded from further analysis, as it only generated a few terms so that a meaningful comparison would be possible, see Table 4.1. Text2Onto was only included in the analysis for 300 abstracts as it was not possible to process all 3066 review article abstracts for "lipoprotein metabolism" listed in PubMed.

We performed a bipartite analysis. We tried to automatically reconstruct the manually created LMO terminology, compared the terms predicted by the four methods to the current LMO terms and also evaluated manually the top 1000 retrieved terms. All automatic comparisons between candidate terms and LMO were not case sensitive.

	Methods									
rank		**TFIDF**		**RelFreq**		**Termine**		**Text2Onto**	**OntoLearn**
1	x	metabolic syndrome		review	x	low-density lipoprotein	x	patient	mutation
2	x	HDL	x	metabolic syndrome	x	cardiovascular disease	x	disease	fish oil
3	x	atherosclerosis	x	diabetes	x	metabolic syndrome		risk	hypercholesterolaemia
4		review	x	atherosclerosis	x	risk factor		effect	serum
5	x	LDL	x	HDL	x	cardiovascular risk		study	progression of atherosclerosis
6	x	cardiovascular disease	x	LDL	x	high-density lipoprotein		level	apheresis
7	x	diabetes	x	cardiovascular disease	x	low-density lipoprotein cholesterol	x	atherosclerosis	omega 3
8	x	dyslipidemia	x	cholesterol	x	high-density lipoprotein cholesterol	x	cholesterol	treatment of hypertriglyceridemia
9	x	high-density lipoprotein		type	x	fatty acid	x	lipoprotein	reductase inhibitor
10	x	cholesterol		article	x	coronary heart disease	x	statin	triglyceride
11	x	low-density lipoprotein	x	fatty acids	x	coronary artery disease		role	adhesion molecule
12	x	cardiovascular risk	x	high-density lipoprotein		clinical trial		syndrome	evolution
13	x	fatty acids		role	x	ldl cholesterol	x	diabetes	purification process
14		article	x	dyslipidemia	x	heart disease	x	trial	prescription omega-3
15	x	insulin resistance	x	low-density lipoprotein	x	diabetes mellitus		protein	omega-6
16		type	x	cardiovascular risk	x	omega-3 fatty acid	x	risk factor	hiv-infected
17	x	statin	x	hypertension		blood pressure	x	treatment	marker of inflammation
18	x	hypertension		combination	x	oxidative stress		event	strong evidence
19	x	inflammation	x	insulin resistance		increased risk		therapy	attractive target
20	x	VLDL		protein		density lipoprotein		review	accelerated atherosclerosis
21	x	lipid metabolism	x	disease	x	cardiovascular risk factor		type	internalization
22		combination		studies		coronary artery		mechanism	scenario
23		role	x	inflammation	x	statin therapy		evidence	protease inhibitor
24	x	oxidative stress		association	x	plant sterol		development	inflammatory cell
25	x	obesity	x	plasma	x	reverse cholesterol transport		use	inflammatory marker

Tab. 4.1: Top 25 predicted terms per method. Listing of the top 25 predictions for TFIDF, RelFreq, Termine, Text2Onto and OntoLearn. Terms relevant to the lipoprotein metabolism domain are marked with x.

LMO								
	Precision				Average Precision			
Top	**TFIDF**	**Termine**	**Text2Onto**	**RelFreq**	**TFIDF**	**Termine**	**Text2Onto**	**RelFreq**
50	**35%**	19%	17%	**35%**	**65%**	54%	38%	54%
200	20%	10%	12%	**22%**	**42%**	28%	23%	37%
1000	**8%**	4%	5%	**8%**	**21%**	12%	12%	20%
LMO + Domain expert								
	Precision				Average Precision			
Top	**TFIDF**	**Termine**	**Text2Onto**	**RelFreq**	**TFIDF**	**Termine**	**Text2Onto**	**RelFreq**
50	**75%**	67%	33%	56%	86%	**89%**	52%	70%
200	**55%**	40%	49%	49%	**74%**	65%	38%	60%
1000	**29%**	20%	14%	28%	**51%**	40%	25%	45%

Tab. 4.2: Precision and Average Precision (rank dependent) for top 50 / 200 / 1000 predictions for 4 methods (TFIDF, Relative Frequency, Termine, Text2Onto) in terms of coverage of LMO and relevant vocabulary. The key finding is that among the top 1000 predictions there are up to 51% terms, which are in the LMO or considered good terms by expert, implying that automated term recognition can play an important role in semi-automated ontology design.

Reconstruction of LMO Terminology

Consider Table 4.2, which shows the percentage of terms that can be generated by the four methods. The first table lists the results for LMO alone, the second for LMO and terms lacking in LMO that were considered relevant after manual inspection. Furthermore, we distinguish precision and average precision. The latter takes the ranking of terms into account:

1.
$$average\ precision = \frac{\sum_{r=1}^{N}(P(r) \times rel(r))}{number\ of\ retrieved\ terms}\ ,\ with$$

2.
$$rel(r) = -\frac{2}{N^2}(r-1) + \frac{2}{N}$$

where r is the rank of retrieval and $P(r)$ is the precision at a cut-off rank. For each of the four methods we list the percentage of relevant terms for the top 50, top 200, and top 1000 predictions. The results show that the precision for the top 50 predictions for LMO ranges from 17-35% and 4-8% for the top 1000 predictions. Using LMO and the expert terms leads to better results of up to 75% for the top 50 predictions and up to 29% for the top 1000. Considering the average precision and thus the ranking of terms, results for the top 50 predictions go up to 89% and for the top 1000 up to 51%. Generally, Termine which favours long terms performs well for the top 50, because long terms are a good indicator of a relevant term. However, there are many short terms, which are relevant, too. The TFIDF and RelFreq methods can pick up these terms, as they include background knowledge, i.e., frequencies of terms in PubMed. By and large, Text2Onto does not perform so well as it neither includes domain-specific background knowledge (as in the case of the TFIDF developed in-house) nor the ranking pursued by Termine, which is biased towards longer frequent terms. Text2Onto suggested short and very general terms, like 'use', 'effect', 'study', 'event', etc. Although we explicitly deactivated the relation extraction part for this experiment, it is not clear why Text2Onto persisted in ranking these terms in the top of the list. Overall, the results are encouraging, as they indicate that a large part of the terminology can be generated automatically with a simple process.

Concerning recall, consider Table 4.3. 3066 documents contain only 53% of the LMO terms literally. TFIDF manages to predict up 39%, which is an encouraging result. Increasing the document base to 50.000 only 71% of the LMO terms are included indicating a possible upper limit. Figure 4.3 provides an overview of the results we acquired from these comparisons. Figures 4.4 and 4.5 provide zoom-ins of Figure 4.3, describing the performance of each method in the top 50 predicted terms.

	LMO terminology predicted by TFIDF		LMO terminology literally contained
	1000	**all**	
300 review abstracts for "lipoprotein metabolism"	8.75%	15.35%	20.98%
3,066 abstracts for "lipoprotein metabolism"	14.99%	38.25%	53%
50,000 abstracts containing "lipoprotein"			71.22%

Tab. 4.3: Coverage of LMO terminology in selected document sets. The table sets the upper limit of terms that can be found with text-mining: Even a large text base with 50,000 documents contains only 71% of LMO terms. TFIDF can predict up to 38% of LMO terms.

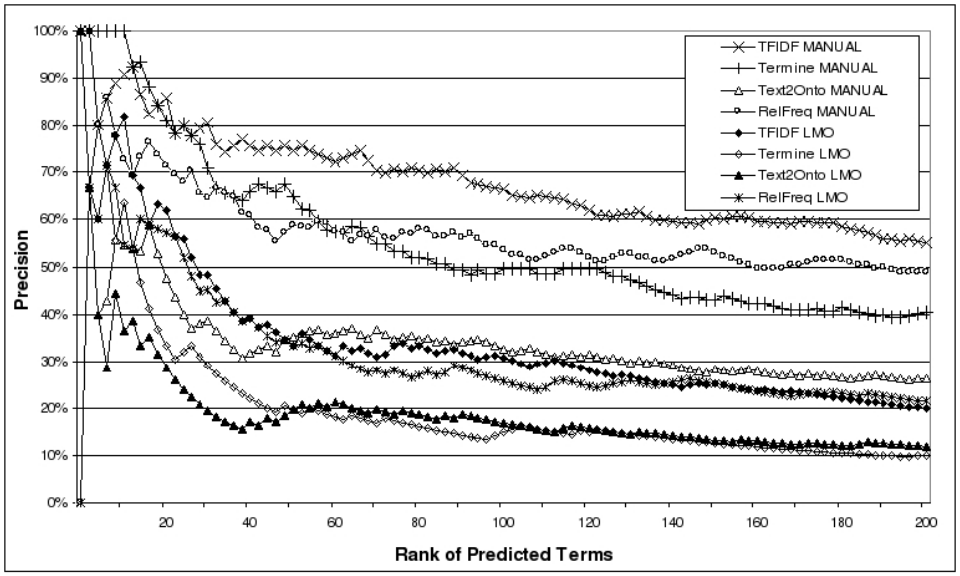

Fig. 4.3: Overlap with manually curated LMO and manual evaluation. Precision at a certain rank r represents each method's capability to recognize domain relevant terms within the top r retrieved terms. The chart shows the overlap within the top r predicted terms with LMO and the manual evaluation (MANUAL). For example, from the top 50 predicted terms by Text2Onto, 20% are in LMO and 36% are correct according to the manual evaluation.

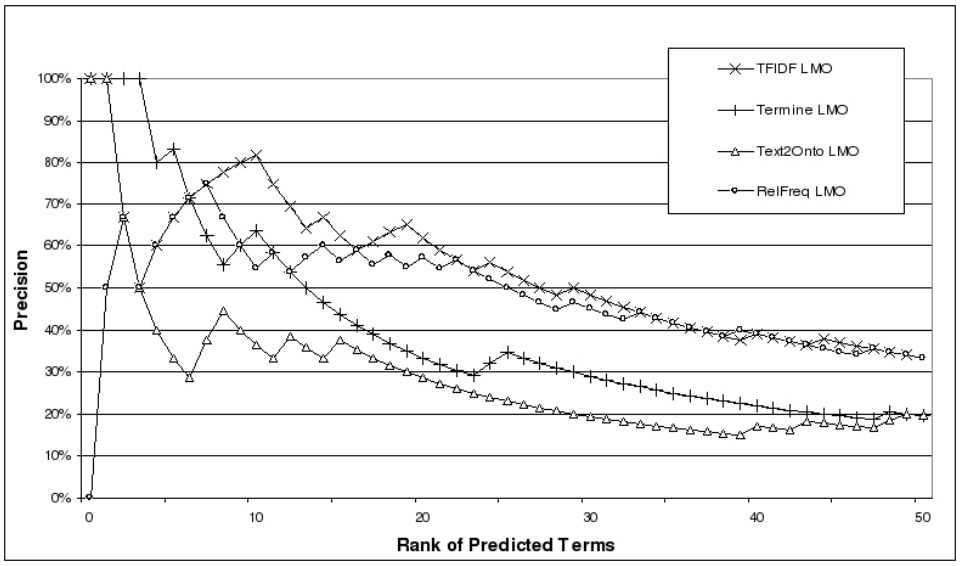

Fig. 4.4: Overlap with LMO. Precision at a certain rank r represents each method's capability to recognize domain relevant terms within the top r retrieved terms. The chart shows the overlap within the top r predicted terms with LMO. For example, from the top 20 predicted terms by TFIDF, 65% are in LMO.

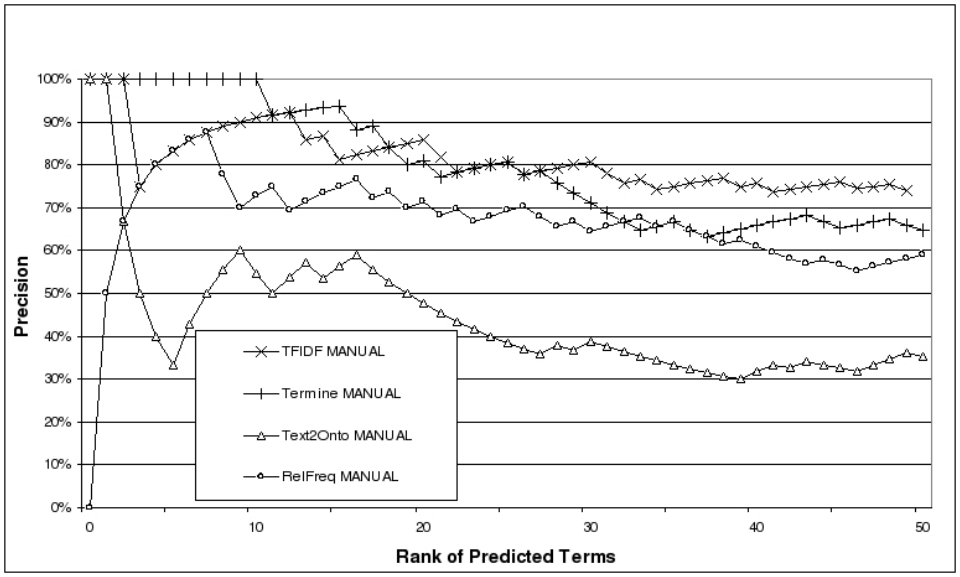

Fig. 4.5: Overlap with controlled lipoprotein metabolism vocabulary and additional manual evaluation (makes sense/makes no sense). Precision at a certain rank r represents each method's capability to recognize domain relevant terms within the top r retrieved terms. The chart shows the overlap within the top r predicted terms with the manual evaluation. For example, from the top 10 predicted terms by Termine, 100% are relevant to lipoprotein metabolism.

4.3 Discussion

The low coverage of the LMO in the data sets calls in question the document set selected and the suitability of the manually built LMO itself. The straightforward approach to select relevant documents from PubMed (review articles in "lipoprotein metabolism") did not return enough documents to cover all of the LMO.

The LMO terms that were absent from the 50,000 PubMed abstracts were grouped in five categories: *rarely occurring terms*, *rarely occurring variants of terms*, *very long terms*, *combinations of terms/variants* and, finally, *terms that should normally be easily found*.

Terms such as 'experimentee' (2) (absolute count of appearance in PubMed per term is given in parenthesis), 'obesive' (2), 'test person' (76) and 'central fatness' (9) are LMO terms, but rarely used by authors and, therefore, rarely appearing in PubMed.

The second group contains variants of terms that appear rarely in PubMed, such as 'Apo-F' (14), 'apolipoprotein c-3' (4), 'IDL I' (1), 'VLDL chol' (34), 'diabetis' (37, instead of 270177 occurrences for 'diabetes'), 'free chol' (0, instead of 2622 for 'free cholesterol'), 'hypolipoproteinaemia' (5, "ae" spelling is rare), 'insuline resistant' (0, instead of 3912 for 'insulin resistant'), 'slo syndrome' (36) and 'sphingomyelinase deficiency disease'(0, MeSH synonym for 'Niemann-Pick Disease'). However, we decided to include such terms in the LMO for completeness.

The third category contains terms that are too long and, therefore, unlikely to appear as such in text: 'receptor-mediated extra-hepatic cellular uptake' (0), 'macrophague cellular uptake' (0), 'predominance of large low-density lipoprotein particles' (0) and 'apob100 containing particles' (2). However, given the initial purpose of the LMO for document retrieval, these terms were included to be recognized by the ontology-based text-mining methods (Doms and Schroeder, 2005).

The fourth group is a combination of the previous two, i.e. LMO terms that are long terms and contain rare variants of LMO terms, such as 'elevated plasma-tg level' (0), 'increased total chol' (0, instead of 116 for 'increased total cholesterol'), 'long-lived test person' (0), 'apoprotein b100 kinetics' (0), 'elevated plasma tg concentrations' (0), and 'decreased hdl-chol' (4).

The last group contains LMO terms that appear often in PubMed and should normally be identified, but are probably absent from the document set, due to its size or specificity. Such terms are 'diabetes type I' (126), 'acetyl-coa c-acyltransferase' (430), 'apolipoprotein-c' (1585), 'type-II diabetic' (1132), 'long-lived population' (23), 'middle-aged adult' (81), 'human body composition' (95), and 'lipid poor HDL' (12).

The third and fourth groups of terms belong to the same category as the hydrolase activity example described earlier in Section 4.1.1. Composite terms like 'receptor-mediated extra-hepatic cellular uptake' and 'predominance of large low-density lipoprotein particles' could be easily broken into several semantic parts (e.g., receptor-mediated/ extra-hepatic/ cellular uptake, or more) and handled by an algorithm that could later compose them and still keep their semantics.

The terms that were predicted by most of the methods but were not in the LMO were further examined and grouped. These were either wrongly predicted ones, meaning phrases frequently occurring in the corpus, but not relevant to LMO, (∼25% of the TFIDF predictions for the Top50 terms) (e.g., 'review', 'type', 'article', 'role', 'event', 'use') or vocabulary that could extend the current ontology (∼40% of the TFIDF predictions for the Top50 terms). This would include disease-specific terms such as 'atherosclerosis', 'cardiovascular risk' and 'atherogenic dyslipidemia', drugs or other chemicals such as 'statins', 'ezetimibe' and 'torcetrapib', or even method and therapy related terms like 'dose' and 'lipid lowering therapy'.

4.4 Conclusion

As pointed out by Castro *et al.* (2006), automated term recognition is missing from many ontology design methologies. In this work, we manually created an ontology for lipoprotein metabolism with 223 concepts and 623 additional synonyms (846 terms in total), we derived design principles and systematically evaluated four methods for automated term recognition.

Automated predictions of up to 1000 terms generate in the order of 40-50% useful terms. Considering only the top 50 terms, the results improve up to 89% *average precision* for LMO + domain expert (defined earlier). This suggests that Automatic Term Recognition (ATR) methods can aid and speed up the process of ontology design by providing lists of useful domain-specific terms, but that they cannot (yet) replace the manually designed term lists. The key problem to further improve these results are composite terms which do not appear literally in text, like GO's 'hydrolase activity, acting on ester bonds' or LMO's 'receptor-mediated extra-hepatic cellular uptake'.

In general, automatic term recognition can have good precision, with up to 75% of the top 50 terms being correct. However, recall is quite low, with 53-71% of terms appearing in text and up to 38% being possible to predict. The ranking doesn't necessarily need to be complex; in our case, a ranking that benefits from the term local frequency in a domain-specific corpus and the term global frequency in PubMed is enough to extract a proper terminology with high accuracy.

Overall, our results show that ontology development can be performed in a semi-automatic way. The domain expert must have as initial input the output from an automatic term recognition method and proceed with enriching the ontology. The experiment as described aims at providing restrictions as well as decision points for including, excluding and reforming ontology terms. Once the domain expert acquires the list of candidate terms, he/she needs to decide on the relations between them. Formulation of questions is one of the most important steps in the ontology design process, helping to step from a list to an ontology.

We discussed principles for development of an ontology with text-mining as intended use, based on our personal experience from the manual development of the Lipoprotein Metabolism Ontology and GoPubMed. We related these principles to the performance of four different ATR methods and their agreement with the manually built LMO. Open problems relate to the choice of suitable text bodies for term recognition as well as generation of composite terms from basic ones.

Chapter 5

Use cases of Word Sense Disambiguation

As already shown in the previous chapters, word sense disambiguation in biomedical context and with the use of biomedical terminologies is a salient issue. With the ever increasing size of scientific literature, finding relevant documents and answering questions has become even more of a challenge. Recently, ontologies have been introduced to annotate genomic data; they can also improve the question answering and the selection of relevant documents in the literature search. Search engines such as GoPubMed[1] (Doms and Schroeder, 2005) use ontological background knowledge to give an overview over large query results and to answer questions. GoPubMed allows users to explore PubMed search results with hierarchical vocabularies such as the Gene Ontology or MeSH.

The GoPubMed infrastructure can be used with any ontology to search for specific scientific literature. Such an example can be vocabularies used in the Edinburgh Mouse Atlas with genes, tissues, and developmental stages of the mouse embryo, a lot of them containing ambiguities.

Semantically-enriched browsing has enhanced the browsing experience by providing contextualised dynamically generated Web content, and quicker access to searches for information. However, adoption of Semantic Web technologies is limited and user perception from the non-IT domain sceptical. Furthermore, little attention has been given to evaluating semantic browsers with real users to demonstrate the enhancements and obtain valuable feedback.

This chapter demonstrates use cases of word sense disambiguation in ontology-based text-mining and more specifically in question answering with the GoPubMed semantic browser and in mouse-anatomy-specific document retrieval (MousePubMed) (published as book chapters, Dietze *et al.* (2008) and Wächter *et al.* (2007), respectively). It also describes a user-centred evaluation framework developed to evaluate Semantic Web Browsers, showing the readiness of common users to exploit the benefits of the semantic web in the life sciences domain. The evaluation framework and results have been presented at the SWAT4LS[2] workshop and published in Oliver *et al.* (2009).

[1] See http://gopubmed.org/

[2] Semantic Web Applications and Tools for Life Sciences, 2008, http://www.swat4ls.org/2008/index.php

5.1 Ontology-based Text Mining

Ontologies and vocabularies such as the Gene Ontology (Ashburner *et al.*, 2000), UMLS (Bodenreider, 2004), MeSH[3] (Nelson *et al.*, 2001), OBO foundry[4] , SNOMED[5] , and GALEN[6] are widely used for annotating biomedical data. They typically contain thousands of terms and cover broad subject areas of biomedical research. Additionally, many species-specific vocabularies for anatomy have been designed covering, among others, plant (Jaiswal *et al.*, 2005), C. elegans (Altun and Hall, 2006), drosophila (Grumbling and Strelets, 2006), mouse (Baldock *et al.*, 2003; Bard *et al.*, 1998), and human (Rosse and Mejino, 2003) anatomy. These vocabularies are used to facilitate communication between scientists in different communities and inter-operability between databases. Annotators, who are usually human, assign terms from such terminologies for example to genes. These assignments are ideally based on direct evidence from literature. Therefore, it is an important problem to automatically identify terms from ontologies in literature to support and even partly automate the annotation process.

However, if terms from ontologies can be found in text, then ontologies can serve directly in literature search. With knowledge-based search engines such as Textpresso (Mueller *et al.*, 2004), XplorMed (Perez-Iratxeta *et al.*, 2003), and GoPubMed (Doms and Schroeder, 2005), the ontological background knowledge can serve to answer questions like the following:

- Which techniques use the Prominin-1 (CD133) marker?
- Which proteins are related to Alzheimer's disease?
- Which hormone is Autistic Disorder associated with?
- Is apoptosis a hot topic?
- Which are leading centers and scientists for liver transplantation?
- Where is the main research done for dengue and leprosy?
- What treatments does the web discuss for Alzheimer?

These types of queries are known as "knowledge queries". The scientific literature and the web hold answers to all of these queries, but it is difficult to obtain them with classical search engines, as they merely present possibly long lists of search results. In contrast, ontology-based search engines can use their hierarchical background knowledge to provide an intelligent filing system, which categorizes results. The categorization gives an overview over large result sets and can be used to answer questions. For example to find the techniques associated with CD133, a query for CD133 will return many documents as a long list in a classical search engine. In contrast, a search engine with ontological background knowledge will identify flow cytometry as a technique and categorize the documents accordingly. The user can then use this hierarchical filing system to select the few articles mentioning techniques and even fewer ones mentioning flow cytometry. Key to this new search paradigm is the background knowledge, which is used to categorize documents. With efforts such as the Gene Ontology (Ashburner *et al.*, 2000) and MeSH(Nelson *et al.*, 2001), the needed knowledge is readily available. MeSH contains for instance the fact that flow cytometry is a technique and the Gene Ontology contains that apoptosis is also known as programmed cell death and that caspases are part of the apoptotic programme.

The central problem of ontology-based search is the mapping of ontology terms to text. As already seen in the previous chapter, the task, known as *term extraction*, is difficult, as authors do not write their abstracts with an ontology in mind. For instance, the mapping must be flexible and map the ontology

[3] nlm.nih.gov/mesh
[4] See http://www.obofoundry.org/
[5] See http://www.ihtsdo.org/
[6] See http://www.opengalen.org/

term "transcription factor binding" to the text "...a transcription that binds...", although it does not appear literally.

Ontologies – Obstacles in Finding Ontology Terms in Text

A fundamental aspect for the work of researchers is the need to share knowledge. In the beginning this was often done without the help of a controlled vocabulary or nomenclature. This is in particular applicable for the biomedical area and life sciences. There are many genes and proteins that have multiple names or identifiers. An example is 'Hnrpa1' which is also known as 'Tis', 'Fli-2', 'heterogeneous nuclear ribonucleoprotein A1', 'helix-destabilizing protein', 'single-strand-binding protein', 'hnRNP core protein A1', 'HDP-1' and 'topoisomerase-inhibitor suppressed'.

Moreover, there seems to be also in some cases a competition for creative gene names like 'Cleopatra', 'Ariadne', 'groucho', 'lost in space', 'brokenheart', 'hairy', 'superman' and many more. There have been efforts to standardize names or at least to reach a consensus for naming. In the context of yeast research and for human genes, for example, there are widely used standards, even if they are not always adhered to in literature. Similar issues arise, if the task is to annotate genes and their function within the categories biomedical process, molecular function and cellular components. One can find that 'cellulose 1,4-beta-cellobiosidase' is also known as 'exoglucanase', 'superoxide-generating NADPH oxidase' as 'cytochrome B-245', 'thiamin' as 'vitamin B1', 'pyrexia' as 'fever', 'heme' can be also found as 'haem', and 'apoptosis' as 'cell death'.

The aim of ontologies is to reduce this problem. They include concepts, synonyms and their relationships. We have already described some of the most popular ontologies in the life sciences in section 2.2.1. A non-trivial aspect is the design and, later on, the evolution of ontologies. With thousands of concepts and definitions, one needs to keep everything intact and all the relations consistent. Discussing consistency, the Gene Ontology follows an informal approach. The transitive closure still has to hold; this means that, if a concept A is_a B and B is_a C then A is_a C has to be true. These inferred redundant relationships are not kept directly in the ontology. This helps to ease the maintenance of the ontology as corrections, modifications and additions only need to check if their direct relations are still valid.

Even though this consistency definition is a pragmatic solution, there are more formal approaches. One such idea is the usage of description logics to formally define concepts and their relations. This was used for instance in the GALEN and SNOMED ontologies (Rector *et al.*, 1996; Spackman, 2004). The advantage of the formal definitions is the chance to automatically check for inconsistencies in the ontology. Supposing that one adds the new fact 'heparin is_a glycosaminoglycan', but it was not yet stated that 'heparin biosynthesis is_a glycosaminoglycan biosynthesis'. Because of the formally defined relations and concepts, this additional relation can be *inferred* with this new fact in the knowledge base.

As already described in Chapter 1, typical problems that arise from mining life scientific literature are *stemming* ('binding' and 'binds' can be reduced to 'bind', but 'organization' to 'organ' not), *missing words* ("...a transcription factor that binds..." should match the ontology term "transcription factor binding"), *complex format of terms* (commas, hyphens, brackets, e.g. 'hydrolase activity, acting on ester bonds'), *synonymity* and *ambiguity* ('development', drug names 'Trial' or 'Act'). The disambiguation methods proposed and tested in Chapter 3, Section 3.3.1 are potential solutions to the latter.

5.1.1 Question Answering with GoPubMed

Traditional keyword based searching gives a long list of results. But finding the relevant documents is only the start; the user has to check whether the results are relevant to him. GoPubMed can answer all the questions mentioned earlier, as it uses the ontological background knowledge, namely the Gene Ontology and MeSH to index search results in PubMed. This allows GoPubMed to categorize the search results, identify relevant terms in the result set and to summarize trends for a topic. This topic can

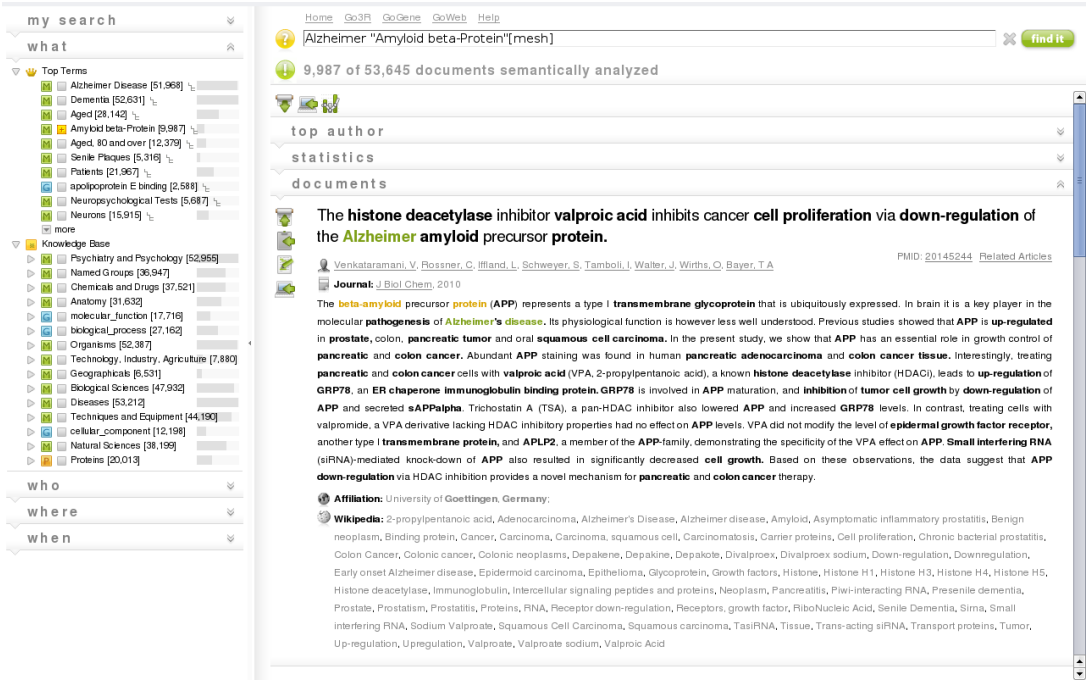

Fig. 5.1: Which proteins are related to Alzheimer's disease? GoPubMed uses its ontological background knowledge to index search results according to the Gene Ontology and MeSH. The interface consists of three parts. The top most part on the right contains the input field for the query, in this example it is "Alzheimer". The left panel contains the ontological background knowledge relevant to the query, put into four answer categories, namely *what, who, where* and *when*. These answer categories contain results such as the representative terminology (what), the most active authors (who) and institutions, places and journals (where), as well as publishing years (when). On the right side, the user can browse the retrieved articles, with the search terms and automatically annotated terms highlighted. Additional links to Wikipedia and proteins identified in text are offered where available. After selecting a term from the left side, here "Amyloid beta-Protein", the result view is updated. It shows now the articles containing the selected concept. This includes also all child terms of the selected term. The initial result set size of 53,645 articles was reduced down to 9,987 relevant articles in two clicks.

either be a term and its children or the result set of a query. For ontology enhanced search in the web, the GoWeb[7] system is available. Figure 5.1 shows a screenshot of GoPubMed. The main panel contains the search results and the panel on the left the relevant categories from the ontologies in a summary and as a tree.

Considering the introductory questions, these can be answered with GoPubMed in the following way[8]:

Question: Which techniques use the Prominin-1 (CD133) marker?
Answer: A search in GoPubMed for "CD133" returns 1,201 documents. Opening Techniques and Equipment in "Knowledge Base" on the left, listed as first is "Flow Cytometry" with 414 articles. The listed articles for flow cytometry contain statements like: " EPC were identified as CD34+/CD133+/kinase insert domain receptor (KDR)+ cells by flow cytometry" (PMID: 20145430). Other interesting terms are "Cell Separation" (115 articles) and "Immunohistochemistry" (111 articles).

Question: Which proteins are related to Alzheimer's disease?
Answer: A search for Alzheimer returns 53,659 documents, 9,989 of which are about "Amyloid beta-Protein" (fourth "Top Term", first protein on the list) By clicking on Amyloid beta-Protein, we can reduce the number or relevant articles and get the following: "A 4-kDa protein, 39-43 amino acids long,

[7] See http://gopubmed.org/web/goweb/
[8] Search results have been acquired on February 2010. They may vary in the future due to the increasing number of publications

expressed by a gene located on chromosome 21. It is the major protein subunit of the vascular and plaque amyloid filaments in individuals with Alzheimer's disease and in aged individuals with trisomy 21 (DOWN SYNDROME). The protein is found predominantly in the nervous system, but there have been reports of its presence in non-neural tissue." The article from Ohyagi Y et al. from 2007 mentions e.g. "Inhibition of aggregation of amyloid pprotein (AP) ... are known as potent therapeutic tools for Alzheimer's disease (AD)." Another article (Chiarini A. et al., Ital J. Anat. Embryol., 2006) states "Reportedly, betaamyloid peptides (Abeta40 and Abeta42) induce the neurodegenerative changes of Alzheimer's disease (AD) ...".

Question: Was Abeta42 already used in a clinical setting?
Answer: A search for "Abeta42 drug clinical trial" into the GoPubMed system retrieves 22 articles. By clicking on "Techniques and Equipment" → "Clinical Trials as Topic", the results are reduced to 6 articles, all of which refer to clinical trials where Abeta42 has been used, as for example in Tarawneh and Holtzman, CNS Neurol. Disord. Drug Targets, 2009, talking about "a recent analysis from a phase I trial that involved active immunization with Abeta42".

Question: Which hormone is Autistic Disorder associated with?
Answer: A search for "Autistic Disorder" in GoPubMed returns 12,677 documents. By selecting "Chemicals and Drugs" → "Hormones, Hormone Substitutes and Hormone Antagonists" → "Hormones", the results are reduced to 537 articles talking about hormones related to Autistic Disorders. Going further down the MeSH hierarchy, one can get results for a certain type of hormone. For example there are only 6 articles on "Estrogens", one of them mentioning that "Estrogen and testosterone have very different effects on calcitriol's metabolism, differences that may explain the striking male/female sex ratios in autism" (Cannell, Med. Hypotheses, 2008).

Hot Topics

Despite the overall growth of literature, some topics are hot and take-off while others are stagnant or are in a cool down phase. Bibliometric analyses aim to shed light on such developments and help to identify emerging trends. Such analyses date back to the 1960s (Price, 1965) and typically focus on research topics (Garfield and Melino, 1997), specific journals (Boyack, 2004), or the researchers themselves (Price, 1965; Newman, 2004). The Hot topic feature of GoPubMed (statistics tab) features views on ontology terms from the knowledge base (see Figure 5.2). It considers a term and all its children as one topic. For each topic a bibliometric analysis is provided.

The hot topic page for an ontology term includes two graphs showing the absolute number of publications per year for a topic (see Fig. 5.3). The second graph shows the relative share compared to the total number of publications per year in PubMed. An increase in the share indicates that the topic is growing faster than the overall number of publications. Both graphs can be used to check whether the publication activity in a topic is decreasing, stagnant, or growing.

Additionally, the user can get a list of the most active authors, the list of journals with the most publications for this topic and a list of cities and countries with the most publications (see Fig. 5.4, 5.5, 5.6). To visualize co-authorship, which author publishes together with which other authors, we provide a co–author network image (see Fig. 5.7). Publications between authors are denoted as edges between the author nodes. If no edge exists then the authors did not yet publish together, according to the publications listed in PubMed for this topic. The last feature is a world map where red dots indicate where all the publications are located for the current topic (Fig. 5.6). All these features of the hot topics page are precalculated using the list of authors and affiliation of an article and the annotations from the GoPubMed system for all 16 Million PubMed articles.
Coming back to the initial questions that need to be answered, we can have the following:

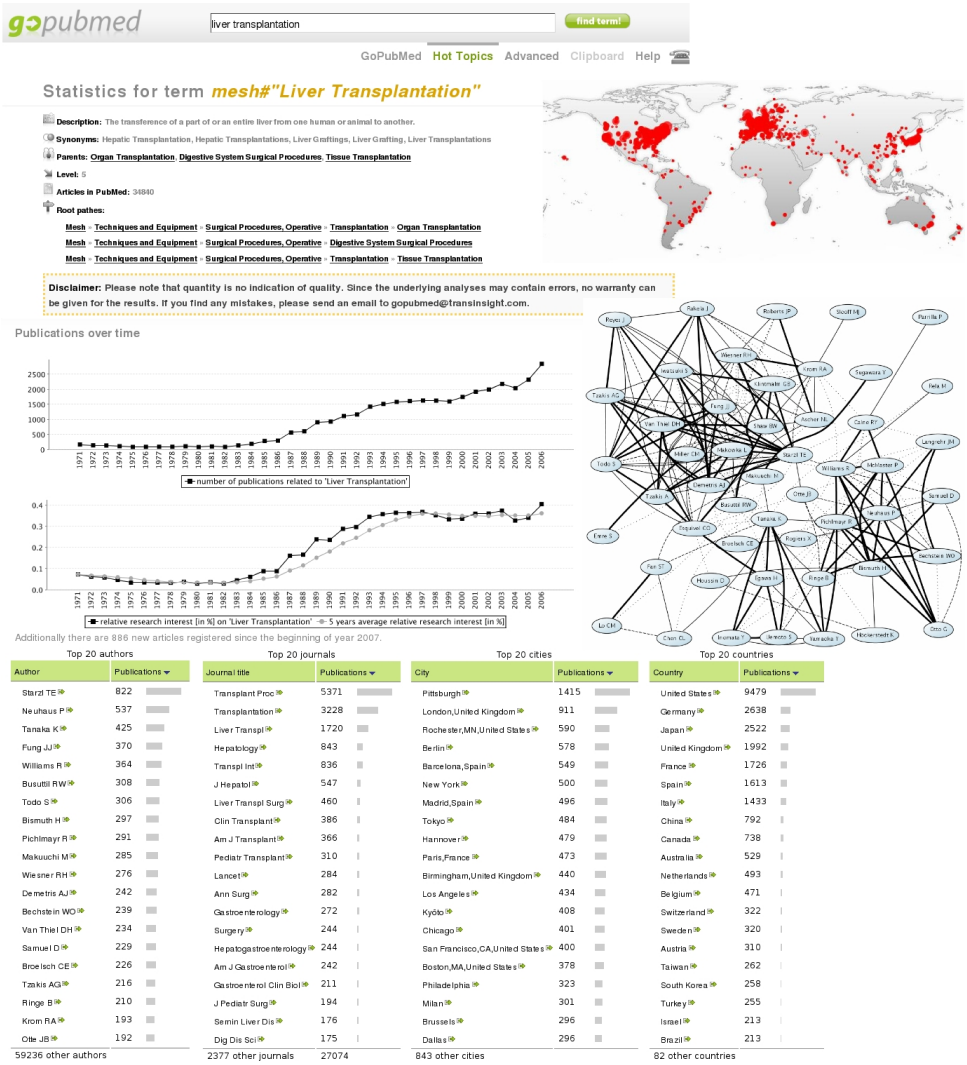

Fig. 5.2: Hot topics in GoPubMed (screenshot of GoPubMed (Doms and Schroeder, 2005)). The result page for a query to Hot Topics starts with a summary for the selected concept including a description, synonyms or the number of all publications in PubMed. Under Publication over time there are two graphs. The first graph displays the number of publications related to this term per year. The second graph visualizes the fraction of publications on the topic over the total number of publications in that year. For "Liver Transplantation" the first graph displays growing number of publication, but the second graph denotes over the last years stagnation in comparison to the overall publication growth in PubMed. The top authors, journals, cities, and countries are presented as tables. All table entries are links and retrieve the related articles. Clicking on "Neuhaus P", for example, retrieves all the publications which have an author with this name. The co-author graph shows which author published together with whom. The thicker an edge is, the more articles contain their names as co-authors. The world map shows the regional distribution of the articles.

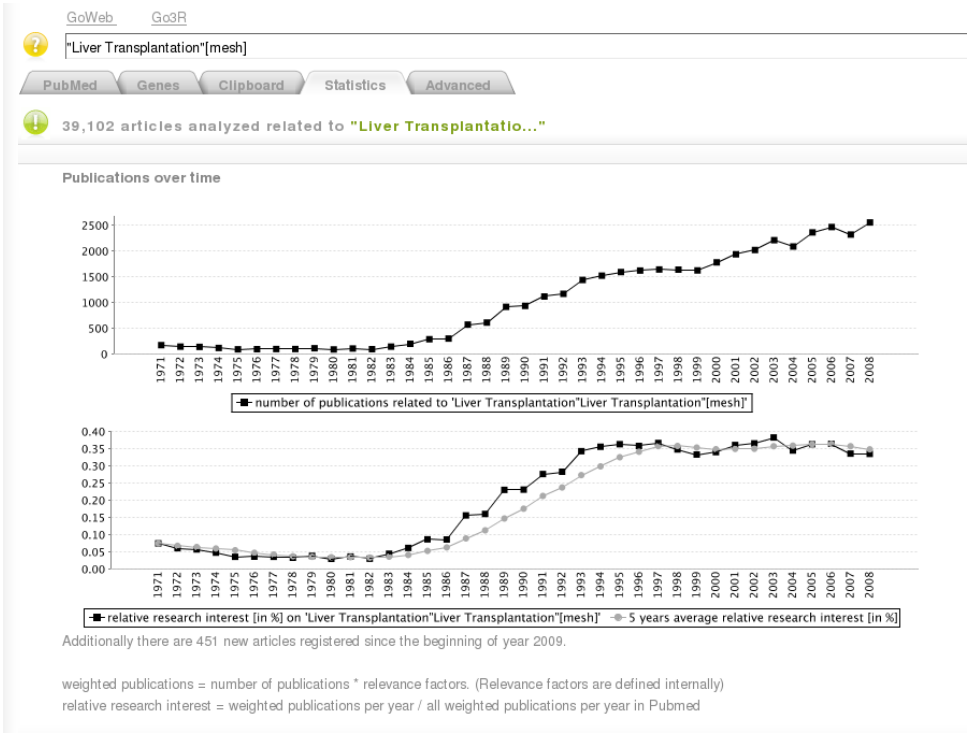

Fig. 5.3: Hot topic page for 'liver transplantation'. The first graph shows the number of publications per year related to the ontology term and the second graph shows the relative share compared to the total number of publications per year in PubMed.

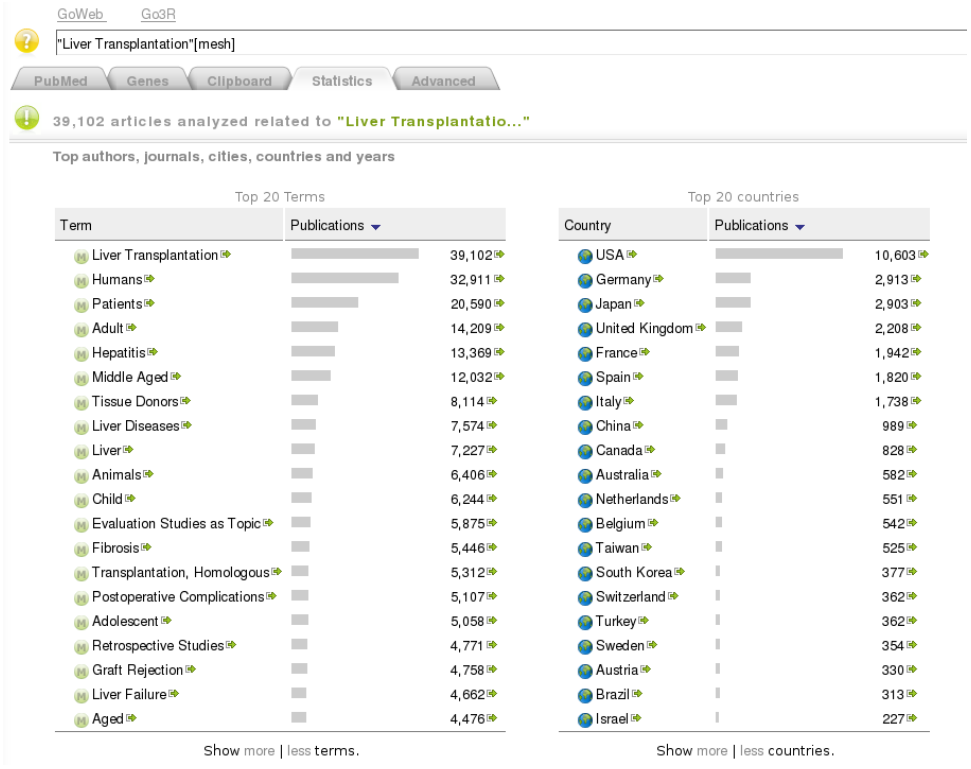

Fig. 5.4: Top co-occurring terms and countries with the most publications on 'liver transplantation'.

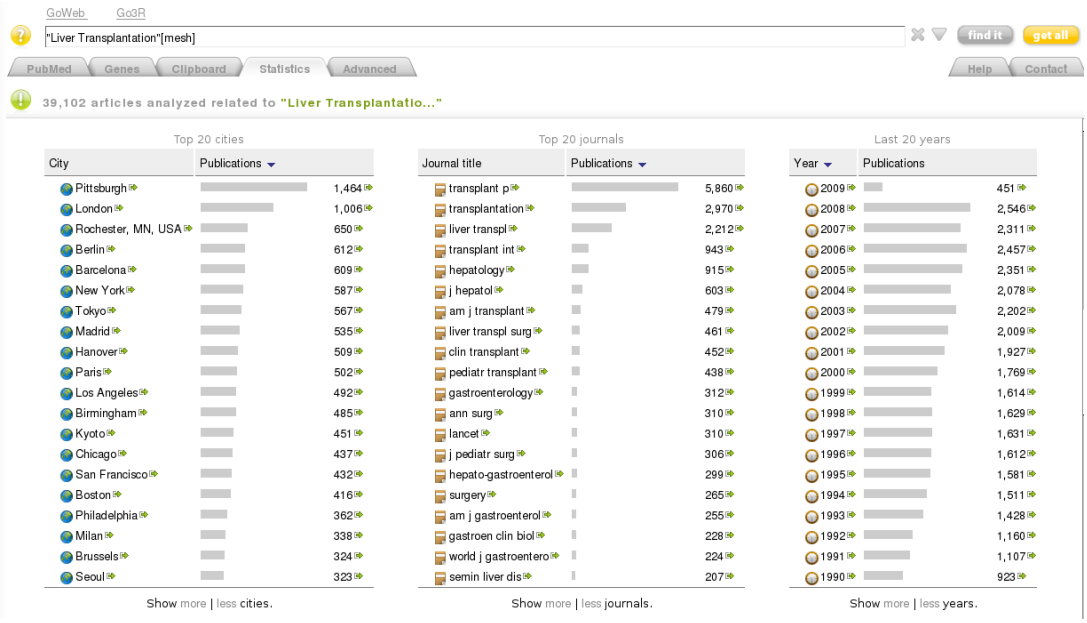

Fig. 5.5: Cities and journals with the most publications on 'liver transplantation' and publication history of the last years.

Fig. 5.6: World map. Red dots indicate locations of all the publications for the current topic.

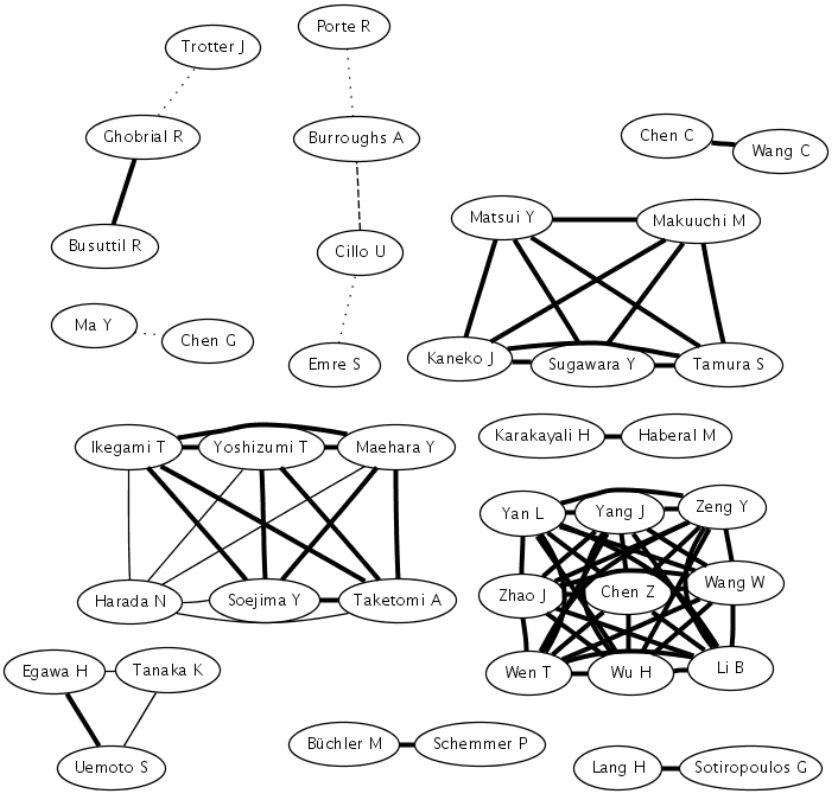

Fig. 5.7: Co–author network. Publications between authors are denoted as edges between the author nodes.

Question: Is apoptosis a hot topic?
Answer: A search for apoptosis and a look at the statistics tab is indicative; the 'publications over time' graphs show that the topic has been growing since the early 1990's. This is in line with Garfield and Melino's (Garfield and Melino, 1997) investigation of the field. But the second graph with the relative research interest shows also, that in the last 3 years the growth was not faster than the average growth of the whole PubMed literature.

Question: Which are leading centers and scientists for liver transplantation?
Answer: A query in GoPubMed for "liver transplantation" and a look at the statistics tab shows that among the top authors is "Neuhaus P" and among the top cities is "Berlin" (see also Fig. 5.2). Prof. Peter Neuhaus works at the Charité Hospital in Berlin, Germany. He is a leading specialist in the field. A look at the coauthor graph reveals with whom Peter Neuhaus has worked and published.

Question: Where is the main research done for dengue and leprosy?
Answer: A search for Dengue and a click on the statiscs tab ends up in a list of top cities and countries for this term. Bangkok and Rio de Janeiro are the two top cities and the top countries are the USA, Brazil, Thailand and India. A similar search for the term Leprosy reveals that India is the top country. This is also reflected in the list of important cities, where one can find several cities located in India. Both terms show that the local occurrence of diseases can be shown in GoPubMed.

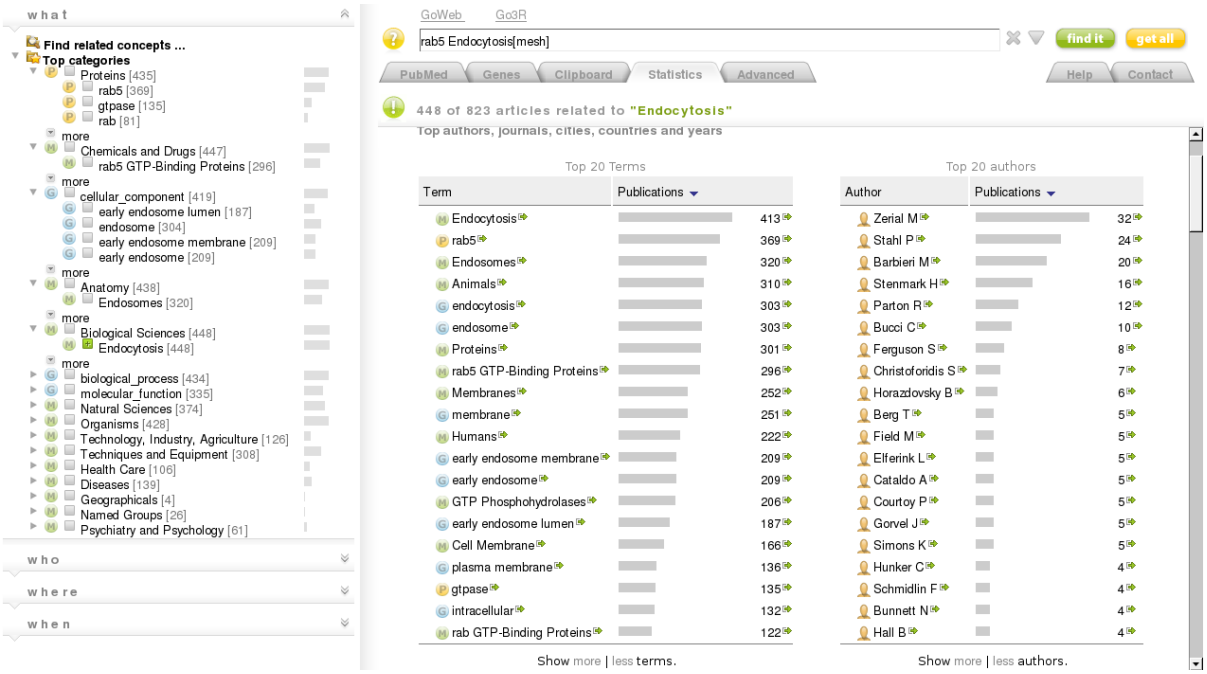

Fig. 5.8: Related topics for Rab5 and 'Endocytosis'. Top co-occurring terms in the Statistics tab are terms like endosome, early endosome, membrane, etc.

GoWeb

Sometimes the search with PubMed is not enough and the user wants to use normal general purpose search engines like Google or Yahoo to search the World Wide Web. GoWeb[9] offers such a global search with ontological background knowledge. Some of the resources the user can search for are for example full text articles not included in PubMed, nonscientific sources like Wikipedia or web based patent databases, commercial sites and vendors for equipment, special interest sites like the Alzheimer Research Forum[10], or even news sites. GoWeb uses standard web search engines and categorizes the results with its annotation algorithms. Normally web searches return not only the url but also the title and a short text snippet from the result page containing the searched keywords. These texts are text-mined and the resulting terms are used in the same way as in GoPubMed to present the results of the search. The user can exploit the ontological background knowledge to answer questions and reduce the result in a fast and efficient way without the need to read all the presented results. It includes, if available, also Wikipedia links and protein names. Coming back to the last of the introductory questions, this can be answered with GoWeb in the following way:

Question: What treatments does the web discuss for Alzheimer?
Answer: A search in GoWeb for "Alzheimer treatment" returns 1,753,277 results. Clicking on the left on "Chemicals and Drugs" reveals results containing the terms 'Memantine' (21 results), 'Amyloid beta-Protein' (6 results), 'Vitamins' (17 results) and others. Clicking on Vitamins reduces the result set from 1,753,277 to 17 documents. In the result snippets the user can find a statement like: "... vitamin may also be an ideal natural treatment for Alzheimer's disease too. ... Over the course of a small study, researchers at the University of Wisconsin ..."

[9] See http://gopubmed.org/web/goweb/
[10] See http://www.alzforum.org/

5.2 Mouse Anatomy Specific Document Retrieval

As already shown in the previous section, ontological background knowledge can serve to answer questions. Consider, for example, a researcher interested in the Pax6 gene. He/she might have the following questions:

- Which processes is Pax6 involved in?
- Which diseases is Pax6 involved in?
- At which developmental stages is Pax6 active in mice?

Literature holds answers to these questions, but a classical literature search cannot answer the questions directly, as articles will not mention gene, disease or process, but rather specific instances such as 'Pax6', 'Aniridia', or 'eye development'. Since ontologies contain knowledge that 'Pax6' is a gene, 'Aniridia' is a disease, and 'eye development' is a process, they can help to answer such questions.

We have already shown in the previous section how ontology-based literature search with GoPubMed can answer questions. In this section, we discuss the use of specialised background knowledge, namely a mouse-anatomy-specific vocabulary built from genes, tissues, and developmental stages as used in the Edinburgh Mouse Atlas (Baldock *et al.*, 2003). We use the GoPubMed infrastructure together with the Edinburgh Mouse Atlas in a system called MousePubMed and we evaluate MousePubMed's automated annotation of PubMed abstracts with the handcurated annotations of the Edinburgh Mouse Atlas. Given emphasis on anatomical and developmental terminology, we discuss the problem of identifying ontology terms in text, with ambiguity being a serious issue. This work has been published as a book chapter in Wächter *et al.* (2007).

5.2.1 GoPubMed and MeshPubMed

GoPubMed (Doms and Schroeder, 2005), MeshPubMed[11] and MousePubMed, which is discussed in the next section, index articles provided by PubMed with ontology terms from GO, MeSH, and Mouse anatomy/development, respectively. As an example consider Figure 5.9, which shows a screenshot of MeshPubMed when queried for Pax6. The key difference to a classical search is that here all the documents are annotated with terms from the domain specific ontology. Therefore, the user interface shows ontological information on the left and the documents on the right side. A list of frequently occurring terms is placed above the complete hierarchy of relevant terms found in documents mentioning the given keywords. Clicking on any of these terms reduces the result set and allows users to quickly filter large result sets to the necessary documents needed to answer their question.

Coming back to the three questions about Pax6 mentioned earlier:

- 'Which processes is Pax6 involved in?' A query in GoPubMed for Pax6 shows that the most frequent process mentioned is *development*. Opening the development branch reveals the processes of *brain* and *eye development* as well as *organ morphogenesis* including *pancreas development*. Indeed the corresponding articles support this "essential role of Pax6 as transcription factor and master control gene in development of eye, brain and pancreas" (Kleinjan *et al.*, 2006).
- 'Which diseases is Pax6 involved in?' A query in MeshPubMed for Pax6 shows that the most frequent disease mentioned is *aniridia*. Hovering the mouse over the term gives an explanation that it is "a congenital abnormality in which there is only a rudimentary iris. This is due to the failure of the optic cup to grow. Aniridia also occurs in a hereditary form, usually autosomal dominant." A click on aniridia shows articles mentioning both the disease and the gene such as for example (Brinckmann *et al.*, 2007), which confirm the answer.

[11] at the time of the experiment the annotation of GO and MeSH terms was separated into two search engines, GoPubMed and MeshPubMed, respectively.

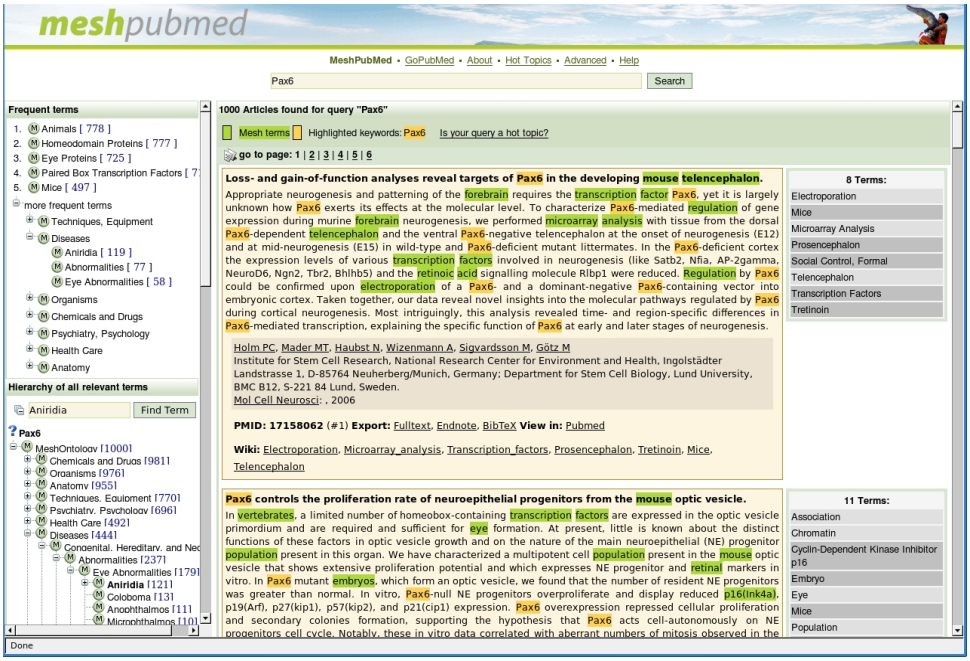

Fig. 5.9: MeshPubMed query for "Pax6". The five most frequent overall terms that occur within the retrieved 1000 abstracts are shown in the upper left corner, together with the most frequent term in the category disease. The hierarchy using the richest path (most abstracts retrieved for certain terms) that starts from disease leads to aniridia (bold, lower left corner).

- 'At which developmental stages is Pax6 active in mice?' A query in MousePubMed for Pax6 shows that Theiler stages up to 14 (9 dpc, days post coitum/conception) are frequently mentioned supporting Pax6's role in early development. Clicking on a stage reveals, e.g., the statement "In the early development of the vertebrate eye, Pax6 is required for..." in Azuma *et al.* (2005).

Indeed, Pax6 is the most researched gene of the family of Pax genes and appears throughout the literature as a 'master control' gene for the development of eyes and is of medical importance because heterozygous mutants produce a wide spectrum of ocular defects such as aniridia in humans.

5.2.2 MousePubMed

To use ontology-based literature search for developmental biology, we built MousePubMed using vocabularies for mouse anatomy (EMAP), human anatomy (EHDA), mouse genes (from EMAGE[12]), and mouse developmental stages (Theiler) as resources from the Edinburgh Mouse Atlas (EMAP[13]) (Baldock *et al.*, 2003). To demonstrate the usefulness of MousePubMed, we evaluate it against tissue and developmental stage annotations in the Edinburgh Mouse Atlas. Before the evaluation, we introduce the matching algorithm developed.

Extracting Gene Names, Anatomy Terms and Developmental Stages

Ontology based text mining is not restricted to finding words or word groups in texts. The structure of the ontology can be used to state the relation between a term and a document by finding the children of the term. This task is reasonably well solvable for the Gene Ontology where its term labels are self-descriptive. Many terms in GO are contained in their child terms (Ogren *et al.*, 2004). As an example, the term "envelope" is refined into "organelle envelope" and further to "organelle envelope lumen". As shown in Table 5.1, anatomical terms can have senses in different domains. These can be common

[12] See http://www.emouseatlas.org/emage/
[13] For Edinburgh Mouse Atlas Project, see http://genex.hgu.mrc.ac.uk/

Term	Other meaning
rod	common English
iris	species: plant; common English
axis	species: deer; common English
chin	common English
beak	common English
pons	protein: Serum paraoxonase/arylesterase 1 (PON)
penis	protein: Penicillinase repressor (penI)
sigma	common English/Greek
patella	species: limpet
cicatrix	disease: scar
nephrons	drug: bronchodilator (Nephron)
hemocytes	drug: iron supplement (Hemocyte)
chondrocytes	drug: cartilage cells for implantation
hippocampus	species: seahorse

Tab. 5.1: Several ambiguous anatomical terms. Some anatomical terms can have other meanings in different domains. Some misinterpretations occur only when certain spelling variations are allowed, for instance, ignored capitalisation or plural forms.

English, drugs, proteins, species, etc. The ontology for the Abstract Mouse contains anatomical concepts in the mouse embryo at different embryonic developmental stages. The vocabulary is used to annotate images of mouse embryos. It unifies the vocabulary needed to describe the different parts throughout 26 Theiler stages. Concepts like organs or body parts are further refined into tissue types, unspecific loci such as "cavities", "left", "upper", as well as general terms such as "node" or "skin". Considering only the textual labels, one cannot distinguish between the different ontological concepts. For example, "chorion" has the children "mesoderm", "ectoderm" and "mesenchyme". "Amnion" and "yolk sac" have children sharing the same labels. Searching for documents related to "chorion" will retrieve very similar document sets to searching for "amnion", only because the documents mention "mesoderm", in this case with meaning "mesoderm specific to amnion". Different anatomical concepts share the same term label. For instance, there exist 171 individuals with label "epithelium". These all refer to different body parts at a specific stage in development.

Ontology-based text mining relies on the assumption that unique or similar types of directed non-cyclic relationships exist, which can be unified in the hierarchical relationships creating a taxonomy. This assumption does not hold for the Abstract Mouse ontology. There does not always exist a path to the common root supported by only one type of hierarchical relationships. Therefore, in our analysis, a document is annotated with a term from the Abstract Mouse ontology taking the term label and its synonymous labels into account. In the Abstract Mouse Ontology the term labels follow various creation patterns. Sometimes a child term contains information of the parent term (for example, "cavities" has the child "amniotic cavity"). In other cases a term like "umbilical vein" has the children "left" and "right", rather than "left umbilical vein" and "right umbilical vein", respectively. These short and common sense labels make the text annotations arbitrary.

For our experiments we slightly adapted the ontology. For the terms "left", "right", "upper", "lower", "common", "anterior" and "posterior" we expanded the term labels with their parents labels. "Eyelids" thus became "upper eyelids" and "lower eyelids", for instance, and we removed the children terms "upper" and "lower" accordingly. To distinguish between common terms such as "skin" occurring for instance, for different organs the matching algorithm took text annotations for ancestor terms into account. Terms with the same label were grouped according to the number of text annotations for their ancestors in the same document. Only annotations of the top ranked group were confirmed. Figure 5.10 shows an example for the term "skin". There were multiple possibilities to resolve this term to a specific tissue. Only when a parental term (shoulder, upper arm, etc.) was found, the text was annotated with

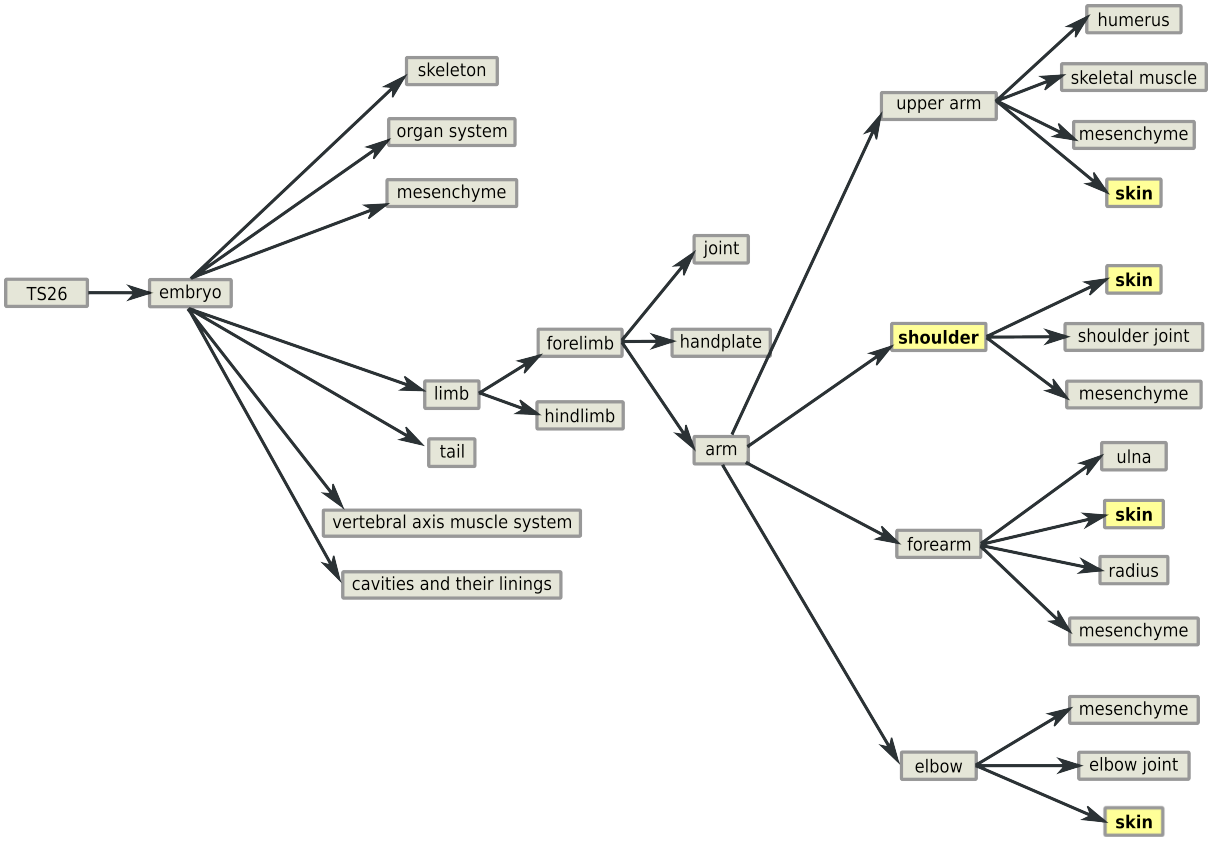

Fig. 5.10: Excerpt from the anatomy ontology, for different types of skin. Occurrences of the term "skin" (yellow concept nodes) in a text were resolved using the hierarchical dependencies. Only when a parental node was also found, for instance, "shoulder", we annotated the text with "skin".

the specific skin.

Finding gene names in documents is done using exact matching against gene names contained in EMAGE. We enriched this set using additional names and synonyms for each gene taken from the MGI database[14]. We tested all 1437 genes mentioned in EMAGE for their annotations with tissues and Theiler stages in PubMed.

We analysed 123,074 abstracts retrieved from PubMed with the query "mouse AND development". This amounted to approximately 0.7 of all documents listed in PubMed. Based on the document annotations with ontology terms, we issued in total 36,358 statements on relations between genes, tissue and developmental stages, which we extracted from EMAP/EMAGE. Cases with multiple Theiler stages from EMAP were split into separate statements. We evaluated the tissues mentioned using EMAP's Abstract Mouse ontology and the anatomy part or MeSH. For path descriptions like "embryo.ectoderm" in EMAP we required the matching document to be annotated with the terms "embryo" and "ectoderm". For MeSH, as in MeshPubMed, we also included descending terms. A document was annotated with the term "embryo" if annotations for its descendants, for example, "germ layers" or its children "ectoderm", "endoderm" or "mesoderm", were found.

To find mentions of Theiler stages in texts, it was not enough to search for them directly, as they seldom occur as such in abstracts ("Theiler stage 12", "TS12", etc.). We therefore compiled a set of regular expressions based on two main notions, the mentioning of embryonic days (E) and of days post coitum (dpc). These expressions had to capture occurrences like

- "embryonic day 10.5",

[14] See http://www.informatics.jax.org.

- "day 9 mouse embryos",
- "between E3.5 (E = embryonic day) and E8.5",
- "12.5 days post coitum", and also
- "7.5-13.5 days post-conception."

As mentionings of Theiler stages do not often occur, but rather general time spans are given ("early embryonic development"), we decided to assign Theiler stages 1 to 14 to "early development", and stages 20 to 27 to "late development," respectively. Every mention of an "early developmental stage" thus was treated as a match for stages 1 through 14. Both assignments were based on statements found in PubMed relating days to general time spans.

Experimental Design

To assess the potential of ontology-based literature searches, we designed two experimental scenarios. For the first, we manually collected two sets of queries and detailed answers. For the second scenario, we evaluated the complete EMAP/EMAGE data. Using the methodology described earlier, we tried to find textual evidences for all sets in PubMed. This means that we searched PubMed for abstracts that shared annotations for each collected triple consisting of a gene, tissue, and Theiler stage.

Manually Curated Test Set

At first we manually selected a set of questions to study the results in detail. The idea was to send simple keyword queries to MousePubMed, asking for mouse abstracts that discuss a certain tissue and embryonic day. MousePubMed should then identify all genes mentioned in the top-ranked abstracts. Questions and retrieved answers were as follows:

- 'Which genes play a role in the development of the nervous system in Theiler stage 14?' A query for "mouse development nervous system 9 dpc" finds the genes Adamts9, Hoxb4, Otx3, and EphA4 within the first eight abstracts. In addition, the genes EphA2, A3, A7, B1, B2, and B4 are found, which are not yet annotated in the EMAGE database.
- 'Which genes play a role in sex differentiation during murine embryo development?' A corresponding query for "mouse sex 10 dpc" results in a set of eight genes within the first fifteen abstracts: Fgf9, Asx11, Sry, Sox9, Usp9x, Maestro/Mro, Wt1, Amh1 and Fra18 . Only half of the genes can be found in EMAGE so far.
- 'Which genes play a role in the development of the murine embryonic liver?' A query for "mouse 'liver development"' results in a set of several genes, most of which can be found in EMAGE as well: Shc, Pxn, Grb2, PEST/Pcnp, GATA6, HNF4a, Foxa1/2, Zhx2, HNF6, Mtf1, SEK1, Nfkb1, c-Jun, Itih-4, and Hex. To answer this question exactly, however, too few abstracts mention particular Theiler stages or days post conception. They rather refer to "early stages of development", and the exact time span might be presented in the full text article only.

All the results, in particular where genes and exact Theiler stages are concerned, are highly dependent on the ordering of abstracts as provided by PubMed. Whenever a new publication appears containing the same search keywords, it will displace abstracts potentially more informative regarding the original question. Abstracts answering the original question might not appear among the first few and be immediately present to the user. However, text mining methods will still extract all the data, even from older publications, and still the right set of articles can easily be found.

The abstracts resulting from a keyword search occur in the same ordering as provided by PubMed. That is, in general, the most recent articles occur first. However, querying for species, tissues, and

Gene	Tissue	Stage	PubMedID
Sparc	retina, RPE, eye	E4.5, E5, E10, E14, E17	**9367648**
Sparc	lens	embryonic day (E)14	**16303962**
Stat3	retina, RPE, eye	-no specific stage-	12634107
Stat3	lens	E10.5	**14978477**
Pedf	RPE	-no specific stage-	7623128
Pedf	retina	E14.5, 18.5	**12447163**
Runx1	inner retina	embryonic day 13.5	**16026391**
Col15a1	conjunctiva, cornea	E10.5-18.5	**14752666**
Otx2	outer retina	-no specific stage-	15978261
Edn1	retina	-no stage-	11413193
IGF-II	eye, cornea, retina, scleral cells	E14	**2560708**
Wnt7b	anterior eye, cornea, optic cup, iris	-no specific stage-	16258938
CDH2	—	-no stage-	9210582
—	lens	-no stage-	9211469
Col9a1	eye, lens vesicle, neural retina, ciliary epithelial cells, cornea	13.5, 16.5-18.5 d.p.c.	**8305707**
Tgfb2	cornea, lens, stroma	-no specific stage-	11784073
Thra	retina	-no specific stage-	9412494
BMP4	retina	E5	**17050724**
Bmp4	optic vesicle, lens	-no specific stage-	15558471
BMP4	lens, optic vesicle	-no specific stage-	9851982
—	eyes	N/A	15902435
Sox1/2	lens	-no stage-	15902435
—	retina, eye axis	E2, E3, E5	15113840
Notch1	eye	-no specific stage-	11731257
Notch2	eye	-no specific stage-	11171333

Tab. 5.2: Expression patterns identified by MousePubMed in articles derived from Thut *et al.* (2001). Often, an abstract does not mention a (specific) developmental stage; —: MousePubMed did not find this particular fact; otherwise: facts as identified by MousePubMed. Given are only tissues related to the murine eye. PubMed identifiers are shown in bold where all three types of information (gene, tissue, developmental stage) are available in the PubMed abstract.

stages still returns the abstracts that discuss the interesting genes. Although corresponding expression patterns might first have been described in older publications, even in recent publications the desired genes reappear quite often.

Reconstructing Outcomes of Large-scale Screening

Thut *et al.* (2001) provided a list of 62 genes found expressed during eye development in mice, together with developmental stage and substructure. Of the 62 genes, 26 were not previously reported (as of 2001); to 16 genes, novel valuable information could be added; 20 genes were fully reported before. Expression patterns were summarised for E12.5, E13.5, E14.5, E16.5, E18.5 and P2. Using MousePubMed, we tried to reconstruct the result of this large-scale screen of 1000 genes.

As Table 5.2 shows, nine PubMed abstracts (in bold) contained the full information as stated by Thut et al., mentioning gene, tissue, and specific stages (days). For most cases, however, not all data were contained in one single abstract. In three cases, we were not able to automatically spot the gene name (left column), in all cases this was due to synonyms lacking in EMAP and MGI. Note that the assessment of recognising genes was based only on genes mentioned in EMAGE. The tissue could be found in almost all of the cases; from most abstracts, even the specific part of the eye could be extracted.

Type of information	Amount of data
Genes with tissues, stages	1437
Genes with at least one non-trivial tissue, stages	1346
Triples of gene, tissue, stage	18,179
Triples of gene, non-trivial tissue, stage	12,782
Tuples of gene, non-trivial tissue	8653

Tab. 5.3: Types of information and quantity contained in EMAGE.

Type of information	Amount of data
Triples of gene, non-trivial tissue, stage	1637 (12.8%)
Tuples of gene, non-trivial tissue	2667 (30.8%)
Genes with at least one tissue and stage	537 (37.4%)

Tab. 5.4: Number of tuples/triples consisting of gene and tissue or gene, tissue and stage found in PubMed abstracts retrieved by the query "mouse AND development".

Complete EMAP Test Set

To evaluate capabilities of automated searches against the complete EMAGE data, the experimental setting was as follows. Genes in EMAGE have annotated tissues, in which they were detected at various stages of embryo development. Thus, we queried MousePubMed with each gene and checked which tissues were mentioned in the resulting PubMed abstracts. This was based on co-occurrence of the gene considering, a tissue, and a Theiler stage (day) in the same abstract. Currently, there are 1437 genes in the EMAGE database annotated with (sometimes multiple) tissues and stages. All in all, we identified 18,179 such triples gene, tissue, and stage in EMAGE. Many of the annotations consist of general annotations for tissue, like "mouse", "embryo", "left", "female", "node". We removed such trivial instances, because they were very frequently found. 12,782 triples referred to specific tissues, and we tried to find these triples using the aforementioned term extraction (also see Table 5.3).

As shown in Table 5.4, we were able to reconstruct 31% of the gene-tissue associations in EMAGE using PubMed abstracts. Only 13% of the full information (gene, tissue, exact stage) was contained in abstracts. All in all, the data recovered from PubMed included information on about 37% of the EMAGE genes. We noted that in many cases, abstracts do not mention specific time points during development. Sometimes, "early" and "late development" are mentioned, which we resolved as described previously in this section. On the other hand, mentions like "in early liver development" could not be resolved to specific overall-stages without background information. Cross-checks revealed that indeed much of the necessary information was only mentioned in the full text of references annotated by EMAP for a specific association.

Conclusion

Ontologies are widely used for annotation. They are also useful for literature search, but the extraction of terms from text is a difficult problem due to the complexity of natural language. Here, we demonstrated the use of the ontology-based literature engines GoPubMed, MeshPubMed, and MousePubMed to answer questions in the context of development. We discussed the specific extraction algorithms needed for MousePubMed and evaluated them small scale on examples relating to eye development and large scale on gene-tissue-stage triple from the Edinburgh Mouse Atlas. We were able to reconstruct 37% of genes, 31% of gene-tissue associations and 13% of gene-tissue-stage associations from PubMed abstracts. These figures are encouraging as only abstracts are used.

5.3 User-centered Evaluation of Semantic Web Browsers

Semantically-enriched browsing has enhanced the browsing experience by providing contextualised dynamically generated Web content, and quicker access to searches for information. However, adoption of Semantic Web technologies is limited and user perception from the non-IT domain sceptical. Furthermore, little attention has been given to evaluating semantic browsers with real users to demonstrate the enhancements and obtain valuable feedback. The EU–funded project Sealife[15] investigates semantic browsing and its application to the life sciences domain. Sealife's main objective is to develop the notion of context-based information integration by extending three existing Semantic Web browsers (SWBs) to link the existing Web to the eScience infrastructure.

This section describes a user-centred evaluation framework that was developed to evaluate the Sealife Semantic Web Browsers that elicited feedback on users' perceptions on ease of use and information findability. Three sources of data: i) web server logs; ii) user questionnaires; and iii) semi-structured interviews were analysed and comparisons made between each browser and a control system. We focus more on the comparison of the semantic browser GoPubMed against PubMed. Details on the evaluation framework and the evaluation of two more semantic web browsers are provided in Appendix B. The evaluation framework and results have been presented at the SWAT4LS[16] workshop and published in Oliver *et al.* (2009).

Semantic Web and Life Sciences As already mentioned in Section 2.3, there is a huge volume of resources available in the web, increasing the difficulty for users to find specific information and make quality judgements (Roy *et al.*, 2006). Especially in the life sciences domain, scientists and medical practitioners need easy access to information about chemical compounds, biological systems, diseases, and the interactions between these entities, which requires this data to be effectively integrated (W3C Interest Group, 2008). The emerging Semantic Web technology (Berners-Lee *et al.*, 2001) aims to provide a solution. Semantic Web technology in the life sciences has the potential to address the urgent needs of clinicians to find specific, quality-assured information under severe pressure of time (Gray and de Lusignan, 1999). Through Semantic Web Browsers (SWBs) using underlying domain ontologies, context-based knowledge integration and semantically enhanced navigation can be achieved. A common assumption in the IT community is that the excitement about the Semantic Web technology will be shared by domain users. However, little attention has been given to evaluating SWBs with real users to demonstrate the enhancements and obtain valuable feedback.

Sealife Project The EU–funded project Sealife (Schroeder *et al.*, 2006) aims at providing easy access to disseminated information and resources in the life sciences' online databases. Its objective is the design and implementation of a semantic Grid browser to link the existing Web to the currently emerging eScience infrastructure. This has been accomplished using eScience's Web/Grid Services and its XML-based standards and ontologies. The main targets of Sealife are the infectious disease and molecular biology domains, illustrated respectively by the National Electronic Library of Infection[17] (NeLI) portal in the United Kingdom, and the National Library of Medicine PubMed[18] publications database (accessible via GoPubMed technology).

To meet the objectives of the Sealife project, browsers have been implemented for different target audiences, including infectious disease clinicians and molecular biologists. As each target group has different needs, prototypes have been developed following the principles of semantic browsing based on structured vocabularies or domain ontologies. To evaluate these distinct browsers, a common evaluation framework was needed.

[15] http://www.biotec.tu-dresden.de/sealife/
[16] Semantic Web Applications and Tools for Life Sciences, 2008, http://www.swat4ls.org/2008/index.php
[17] See http://www.neli.org.uk/
[18] See http://www.ncbi.nlm.nih.gov/pubmed/

In this section we outline the work conducted to design a common evaluation framework for the Sealife Semantic Web Browsers and the hypotheses that were tested. While a Web browser navigates along links between documents, a SWB navigates along relationships in a web of concepts (Berners-Lee *et al.*, 2007). We use the term *Semantic Web Browser* (SWB) for any browser which:

- uses at least one knowledge organisation system (KOS), either a structured vocabulary or an ontology, to support the browsing;
- is able to identify and highlight "useful" terms in the content being visited;
- enables semantic interpretation of these Web pages and adds semantic hyperlinks to their highlighted terms,
- gathers additional information from the highlighted terms, which may involve access to external data and services (e.g., European Bioinformatics Institute or PubMed) (Diallo *et al.*, 2008b) called *targets*.

This is the first user-centred evaluation of SWBs to be conducted using established, real-world Web resources as control platforms and recruiting participants from among the real-world users of these resources.

The Sealife Semantic Web Browsers

To make the evaluation framework more comprehensible, we describe briefly in this section the different implementations of the 3 Sealife browsers. The first browser, COHSE-NeLI, is based on the Conceptual Open Hypermedia System (COHSE) (Yesilada *et al.*, 2008) developed by the University of Southampton and the University of Manchester. The second is the CORESE-NeLI framework (Diallo *et al.*, 2008b) based on the CORESE engine developed at INRIA. Finally, the GoPubMed/GoGene SWBs are developed at the Technical University of Dresden (Doms and Schroeder, 2005).

The COHSE-NeLI SWB The COHSE system (Yesilada *et al.*, 2008) automatically adds hyperlinks on Web pages by recognising and highlighting terms contained in background knowledge, based on an ontology or KOS (Figure 5.11). When a highlighted term is clicked, a link box appears (see Figure 5.12), populated with links to trusted external resources. For any highlighted term, resources are provided for broader, narrower, and related terms (e.g. affects/is_affected_by, is_symptom_of/has_symptom, causes/is_caused_by, treats/is_treated_by) obtained from the vocabulary. For the Sealife project, COHSE was adapted for the NeLI portal, and the version discussed here uses the NeLI vocabulary (Diallo *et al.*, 2008a) enriched with MeSH terms (Nelson *et al.*, 2001) as its KOS. The NeLI vocabulary formalises the infectious disease domain and is modelled in the SKOS language[19].

The CORESE-NeLI SWB The CORESE-NeLI (Diallo *et al.*, 2008b) engine supports the navigation of a portal by the use of a knowledge artefact (either a structured vocabulary or a domain ontology). The browser can perform a) a semantic search and b) semantic browsing of the NeLI portal. The CORESE-NeLI engine bases its semantic search on semantic annotations generated from Web pages using the NeLI vocabulary, and using the relationships in the knowledge artefact (i.e., narrower, broader, related to) to retrieve annotated pages related to the user query. For semantic browsing, CORESE-NeLI can identify and highlight, in a Web page being visited, terms retrieved from a structured vocabulary. From the highlighted terms, it can then create links to related pages within the portal, enabling semantic browsing. A query can be built from the highlighted terms to query external resources such as Google and PubMed.

[19] See http://www.w3.org/TR/skos-reference/

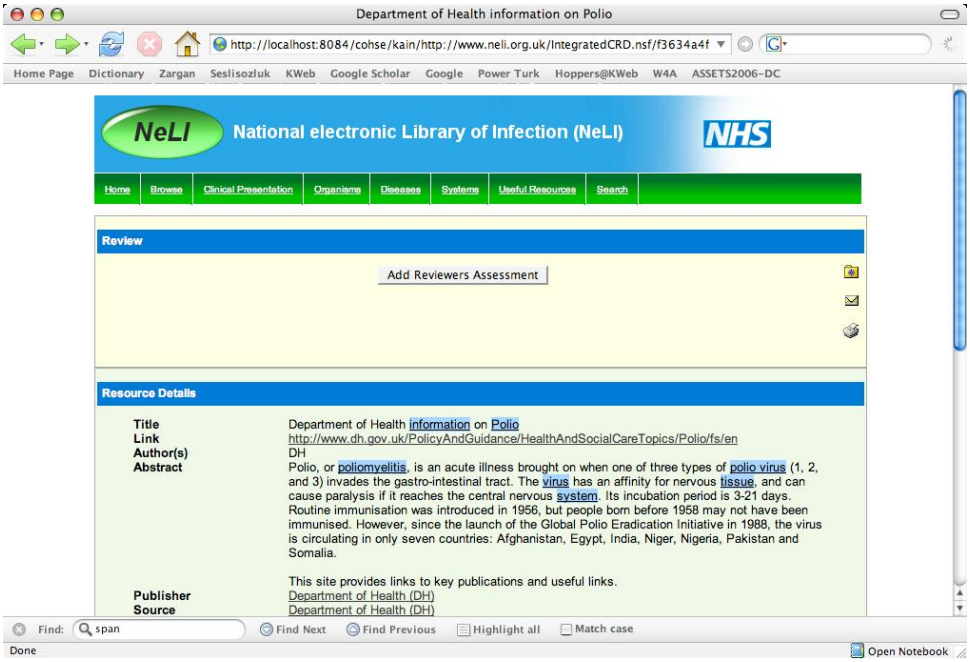

Fig. 5.11: COHSE semantic links as seen on the NeLI portal.

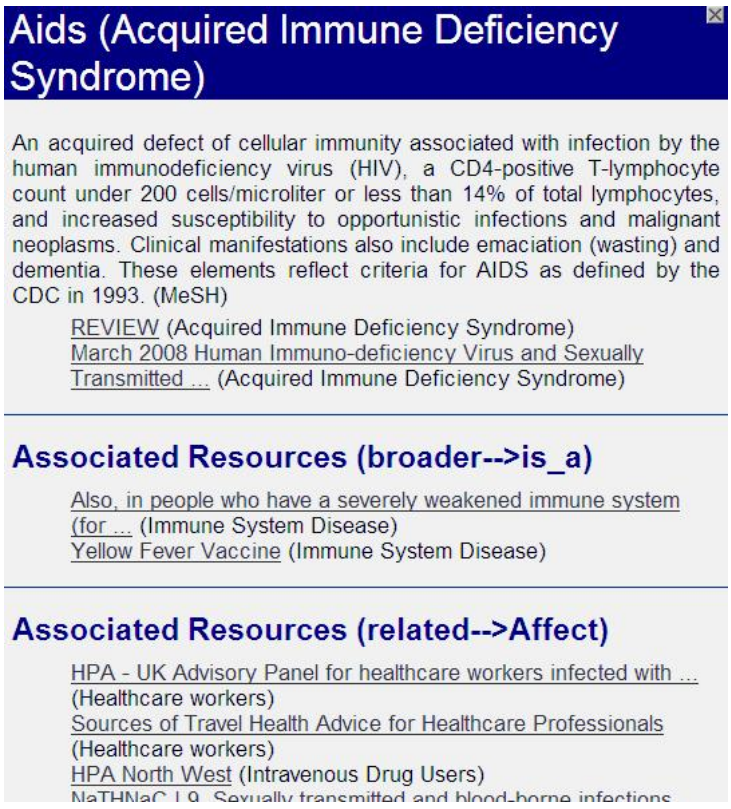

Fig. 5.12: COHSE semantic links: link boxes which appear after a click on the highlighted terms.

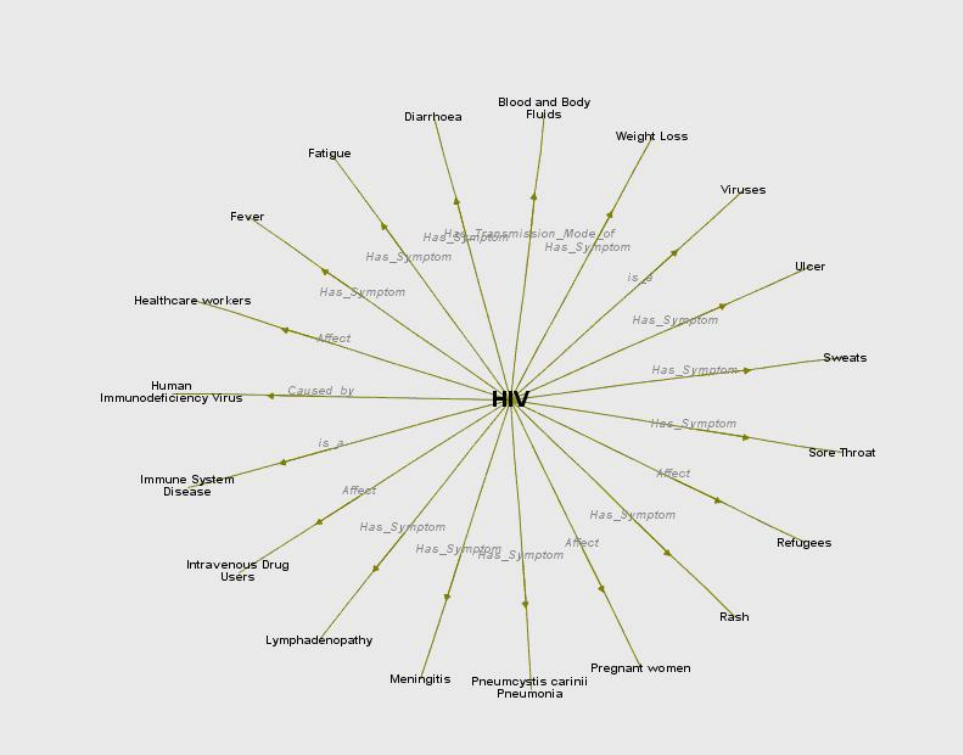

Fig. 5.13: The CORESE search box and graph showing terms related to "HIV" (Human Immunodeficiency Virus).

CORESE is accessible via a plugin in Firefox. Entering a search term which exists in the NeLI vocabulary opens a tabbed pane. The first tab shows a graph of related concepts in the NeLI vocabulary (Figure 5.13), with which the user can navigate the NeLI Digital Library (DL) by double-clicking on a node or an edge. To the left of the graph is a history of recently visited search terms. The second tab shows a list of related documents (Figure 5.14).

GoPubMed and GoGene for molecular biology GoPubMed and GoGene are search technologies applied to the PubMed online database. GoPubMed, previously described in Section 5.1, uses ontologies to deal with the wealth of medical and biological research literature by grouping literature by the underlying information in the abstract. GoPubMed offers name recognition and computational Web

Fig. 5.14: CORESE-NeLI pane of related documents.

services. One of the major problems in text mining is the ambiguity of names of genes and proteins (especially crucial for computational Web services), as well as context-based terms used in molecular biology. GoPubMed and its underlying search engine handles this problem. GoGene[20] associates all genes from different model organisms with concepts of GO and MeSH. The hierarchical structure of the vocabularies allows long lists of genes to be clustered and summarized. Because most knowledge is contained only in publications and not in databases, GoGene integrates manually curated gene annotations, literature references (GeneRIFs) and textual comments from UniProt and EntrezGene with text-mined annotations from all abstracts in PubMed. In doing so, more than 4,000,000 associations between genes and ontology concepts for the model organisms human, mouse, rat, worm, fruitfly, zebrafish, thale cress, baker's yeast, fission yeast, and E. coli are made available, thereby increasing the number of known GO annotations by one order of magnitude. Additionally, GoGene provides 35,000 gene-mutation associations extracted from PubMed abstracts, which are not contained in UniProt. All associations are linked to their origin (i.e. literature or database entries) for further investigation. By scanning the literature, GoGene also compiles publication histories for each gene. Such histories allow to rank genes according to what is new (many new publications recently), what is widely studied (many contributing authors), or what is of high impact (accumulated journal impact points). All relevant concepts for a gene list are displayed as a tree that allows quick navigation through long lists of genes.

5.3.1 Aims and Objectives

In the Semantic Web area in general, some comparable evaluations raising interesting issues have already been reported in the literature (Uren *et al.*, 2005; Reichert *et al.*, 2005). The EON workshops initiative (International Workshops on Evaluation of Ontology-based tools) provides an environment for technical evaluation of Semantic Web tools. The aim of this user-centred evaluation of Semantic Web Browsers was to compare each SWB not to the other SWBs, but to a *non-semantic* control platform.

The following hypotheses were made to test the key purposes of the SWBs: improving mobility and travel within the system and improving user satisfaction.

Mobility and travel within the system

- **H1**: The SWB reduces the time taken for users to find information or perform tasks.
- **H2**: The SWB shortens the pathway taken to find information or perform tasks.
- **H3**: Where semantic links are available, users will always follow them instead of non-semantic links.

User attitude and satisfaction

- **H4**: Users find the SWB easier to use than the control platform.
- **H5**: Where semantic links and ranking are available, users prefer them to non-semantic links and ranking.
- **H6**: Use of the SWB is intuitive:
  - **a)** Users think the SWB helps them to find information or complete tasks.
  - **b)** Users intuitively understand how to use the SWB to find such information or complete tasks.

An evaluation framework was then designed to test the above hypotheses. More details on the framework are given in Appendix B and in Oliver *et al.* (2009). In the following Section we describe one part of the evaluation, which is the comparison between the Semantic Web Browser GoPubMed and the control system PubMed.

[20] See http://projects.biotec.tu-dresden.de/gogene/gogene/

5.3.2 GoPubMed vs. PubMed – Results

The evaluation of GoPubMed took place in the form of a workshop with 20 students from the Masters Program in Molecular Bioengineering at the BIOTEC, TU Dresden. The students were first given a short introduction to GoPubMed/GoGene functionality. They later had to fill in a pre-questionnaire concerning their scientific background and their way of searching for scientific literature. They were then given a series of tasks to perform with the use of PubMed, GoPubMed and GoGene (see following subsection tasks – main test). The workshop finished with the students answering a post-questionnaire comparing the unmodified system (PubMed) to the modified one (GoPubMed).

Pre-questionnaire

Based on the filled-in pre-questionnaires, 45% of the participants had a main University degree in Biology, 25% in Engineering and 30% in other fields. Their experience in this field was for the 65% under 5 years and for the 35% 5 years or more. Only 6.25% of the participants use PubMed daily, 31.25% more than once a week but not as often as daily, 12.5% once a week and the 50% of the participants use PubMed for more than once a month but not as often as once a week. On average, they rate the usefulness of PubMed to 68.75% Concerning what they search for in PubMed, the participants have chosen between the following answers (being able to choose more than one):

75% I look for specific articles

75% I search for reviews for an overview

19% I look for papers in a specific journal

50% I look for papers of specific authors

62.5% I look for papers for specific diseases, genes, etc.

31% I look for the most recent papers.

Concerning the use of other search engines for their research, the participants have chosen between the following answers (being able to choose more than one):

94% PubMed[21]

31% Google Scholar[22]

100% Google or other web search

6% Scopus[23]

19% Other specialist

[21] See http://www.ncbi.nlm.nih.gov/pubmed/
[22] See http://scholar.google.com/
[23] See http://www.info.scopus.com/

Tasks – Main Test

The participants were given a set of questions they should answer with the use of PubMed and a set of similar questions to be answered with the use of GoPubMed. These were especially conceived so as not to favour one of the two systems (PubMed vs. GoPubMed) and to avoid as much as possible the bias for a positive opinion towards GoPubMed. The participants were told not to spend more than 8 minutes to answer each question. The participants were divided into two groups, answering half of the questions with **PubMed** and the rest with **GoPubMed** and vice-versa. The two groups of questions were the following:

Group A

1. Which particular diseases are associated most often with HIV?
2. What kinds of diseases are also related to HIV?
3. Which techniques of treatment are used to help HIV patients?
4. Who are the top authors for Antiretroviral Therapy?
5. Where was this research done by those authors?
6. Which are leading centres for liver transplantation?
7. Which are leading scientists for liver transplantation?
8. Is the research on leukaemia decreasing?
9. Which proteins are related to Alzheimer's disease?
10. What is the role of MMS2 in cancer?

Group B

11. How does BARD1 regulate BRCA1 activity?
12. Which are the different types of Paget's disease (where does it locate)?
13. How can Paget's disease be diagnosed?
14. Is there a treatment/therapy for Paget's disease?
15. How rare/prevalent is this disease?
16. Which sex and age groups are the most affected?
17. Which are the leading 3 countries doing research on Paget's disease?
18. Is there anybody in Brazil doing research on Paget's disease?
19. Which are the top Brazilian authors for Paget's disease?

		PubMed	GoPubMed
1	I would like to use this system frequently	5.1	8
2	The system was easy to use	4.9	7.8
3	The system was too complex	5.3	4.3
4	The user interface was easy to understand	6.8	7.7
5	The system responded fast	7.3	6.3
6	The system provided enough help information and examples	3.5	6.8
7	Finding the answers to the tasks with the system was easy	3.8	7.6
8	Finding the answers to the tasks with the system was fast	3.8	7
9	A lot of information the system found was irrelevant to the tasks	7.5	4
10	Most of the information returned by the system was relevant	3.7	7
11	The amount of relevant information I found was less than expected	4.9	3.4
12	The amount of relevant information I found was same as expected	4.3	5
13	The amount of relevant information I found was more than expected	5	5.6
14	The modifications were mostly relevant to me	–	6.8

Tab. 5.5: Post-questionnaire on GoPubMed vs. PubMed. The numbers for PubMed and GoPubMed are averages for agreement on a 1 to 10 scale (1 strongly disagree, 10 strongly agree).

Post-questionnaire

The participants were given a post-questionnaire after completing the tasks and were asked to fill it on a 1 to 10 scale (1 strongly disagree, 10 strongly agree). The results are shown in Table 5.5.

The post-questionnaire included also the following questions, giving the participants the freedom to add more comments and suggestions:

15. Did you find the highlighting of ontology terms helpful?
16. Did you get an overview over your search results from the tree on the left?
17. Did you manage to navigate efficiently through the tree?
18. Did you find any papers you would probably have missed with PubMed?
19. What do you like/dislike about using the tree to explore your search results?

Most of the comments from the participants concerned the appearance of the browser, e.g. some could not find an obvious link to the original paper site (could not easily locate the PMID link in light grey). Others were asking for functionality such as information on how often the article has been cited or read. All of the participants were very positive towards GoPubMed and GoGene, but still had concerns, since 94% of them have been using PubMed and were used to its simple interface.

5.3.3 Conclusion

For GoPubMed a number of the hypotheses formulated at the start of the evaluation were confirmed, especially regarding ease of use. For the other two Semantic Web Browsers, COHSE-NeLI and CORESE-NeLI, most of the hypotheses were contradicted. Table 5.6 shows how user feedback from each system agreed or disagreed with the hypotheses.

The evaluation study demonstrated that the evaluation framework is suitable for eliciting user perceptions of SWBs. The results have allowed us to answer our initial hypotheses fully for each SWB even

	Hypothesis	COHSE	CORESE	GoPubMed
H1	The SWB reduces the time taken for users to find information or perform tasks.	No	Yes	No
H2	The SWB shortens the pathway taken to find information or perform tasks.	No (targets not found)	No (targets found by few users)	PubMed data not available for comparison
H3	Where semantic links are available, users will always follow them instead of nonsemantic links.	No	Yes	No
H4	Users find the SWB easier to use than the control platform.	Yes and No	No	Yes
H5	Where semantic links and ranking are available, users prefer them to nonsemantic links and ranking.	Yes	Yes	Yes
H6	Use of the SWB is intuitive: **a)** Users think the SWB helps them to find information or complete tasks.	No	No	Yes
	b) Users intuitively understand how to use the SWB to find such information or complete tasks.	No	No	Yes

Tab. 5.6: Confirmation or contradiction of original hypotheses.

though each SWB had a distinct implementation and used different aspects of the SW technology. A new evaluation framework for SWBs was designed and tested on 3 intervention Semantic Web Browsers, with participants recruited from the intervention systems' real-world target audiences. The control platforms were live, real-world systems with substantial numbers of existing users. Using this evaluation framework, all of the initial hypotheses were successfully confirmed or contradicted (Table 5.6).

Overall, the framework successfully elicited a range of feedback on 3 distinct Semantic Web technologies. It was found that, although potentially easier to elicit feedback via online questionnaires, observing respondents in a workshop setting provides an excellent opportunity to gather both quantitative and qualitative data from larger numbers of users.

The evaluation showed that users tended to prefer the system (GoPubMed) that had the most mature interface, but were able to use the semantic features of all systems regardless of the interface or types of semantic links presented. The evaluation feedback will contribute directly to future versions of each Semantic Web Browser and there will be further analysis of the weblogs to determine the specific types of semantic links that were or were not used.

Chapter 6

Summary and Future Work

6.1 Open problem 1 revisited

> **Open problem 1:** Word sense disambiguation (WSD) is required for the accurate analysis of text in many applications. Since 2004, the most active domain-specific application area for WSD seems to be bioinformatics (Liu *et al.*, 2004; Schuemie *et al.*, 2005; Edmonds and Agirre, 2006). Classical approaches to WSD use co-occurring words or terms. However, most treat ontologies as simple terminologies, without making use of the ontology structure or the semantic similarity between terms.

We have addressed this problem in Chapter 3, where we used co-occurrences (*Term Cooc*, Sections 3.2, 3.3.1), document clustering (see Section 3.2), the ontology structure (*Inferred Cooc*, Section 3.3.1) and semantic similarity between terms (*Closest Sense*, Section 3.3.1), as well as metadata like the year of publication, journal and abstract title (*MetaData*, Section 3.3.1) in order to perform disambiguation of terms in abstracts of biomedical publications. We furthermore made available a corpus of 2600 documents divided into three datasets of varying quality and quantity that can be used as benchmarks for disambiguation.

The comparison of the methods shows that metadata and training data of high quality are key points for increasing the performance of disambiguation, with up to 96% accuracy (*MetaData* method, trained on high quality/low quantity dataset). However, the production of high quality training data is a tedious and time-consuming process. When such training data are not available, the co–occurrence of ontology/taxonomy terms can be used for disambiguation with high accuracy. The hierarchical structure of the ontology can also improve the accuracy, especially when the ontology is consistently modelled. In Section 3.3 we have showed that a 'is_a' hierachy like the Gene Ontology gives higher disambiguation accuracy compared to a 'narrower_than' hierarchy such as the Medical Subject Headings.

For disambiguation one has to balance between achieving high accuracy and producing training data of sufficient quality and quantity. The MetaData method gave the best results but it required high quality training data, which were hard to produce. The Term Cooc and Closest Sense methods gave lower accuracy than the MetaData. However, they are semi-automated, requiring no manual intervention for training.

Future work on disambiguation

Future work can include several aspects ranging from the use of negative co–occurrences, disambiguation in full-text articles, to a combination of the three methods (Term Cooc, MetaData, Closest Sense) and a decision based on a confidence score for each of the approaches.

While performing disambiguation with the Term Cooc and Closest Sense methods, all terms found in the abstract apart from the term in question were considered as true with respect to the ontology. However, they could as well be ambiguous terms and therefore insert an error into the disambiguation process. In the future, we want to take such ambiguous terms into account as well.

A possible extension could be to correctly identify if a sense occurs that is not included in the ontology and possibly add it. This can potentially be done by setting a threshold. In the Closest Sense approach, from all distances below that threshold, one would be clearly shortest. If not, then this would be the new sense. For the Term Cooc and MetaData methods this could be done by training each method on each sense and if the sense found would be below the threshold, this would indicate a new one.

Another interesting aspect can be the automatic identification of an ambiguous term. So far, the ambiguous terms tested were empirically identified. A more thorough and automated identification pipeline employing WordNet, noun phrase statistics and expert input could be set up.

It would also be interesting to see how the accuracy would change once the disambiguation would be performed in the full text of articles. Co–occurrences could also be computed based on the full-text instead of the abstracts of articles. The number of terms occurring in a document could also be considered in the disambiguation pipeline. We have noticed that in most of the cases where the ambiguous term had one of the false senses, it usually co–occurred with only a few other terms (or in a lot of cases it was the only term in the document).

WSD use cases

In Chapter 5 we demonstrated use cases of word sense disambiguation in ontology-based text-mining and described a user-centred evaluation framework developed to evaluate Semantic Web Browsers. As presented in Section 5.1, the GoPubMed infrastructure can be used with any ontology to search for specific scientific literature. An example of such a search was the mouse-anatomy-specific document retrieval presented in Section 5.2, where genes, tissues, and developmental stages of the mouse embryo contained many ambiguities. We additionally described a user-centred evaluation framework developed to evaluate Semantic Web Browsers in Section 5.3, where we mainly focused on the user satisfaction about GoPubMed.

6.2 Open problem 2 revisited

> **Open problem 2:** Which are the common obstacles during the design of an ontology to be used for text mining? Can automatic term recognition (ATR) methods assist the ontology generation process?

In Chapter 4 we presented the experience acquired during the manual development of a lipoprotein metabolism ontology (LMO) that was afterwards used for text-mining. We manually created an ontology for lipoprotein metabolism with 846 terms in total, we derived design principles and systematically evaluated four methods for Automated Term Recognition (ATR).

We have shown that automated predictions of up to 1000 terms generate in the order of 40-50% useful terms. Considering only the top 50 terms generated, the results improve up to 89% average precision for terms that make sense to be included into the terminology (*LMO + domain expert*[1]).

Based on the results of the comparison between the manually built terminology and the terminologies extracted from the automatic term recognition methods, we have shown that ATR methods can provide lists of useful domain-specific terms. In this way, ATR methods can aid and speed up the ontology

[1]Some terms were not included in the manually created terminology because the domain expert missed them. However, these made sense to be included into the LMO, therefore the *LMO + domain expert* set contains manual terminology together with automated terminology that was domain related

design process, but the terminologies produced cannot yet replace the manually created terminologies, nor construct ontologies without any expert intervention.

Composite terms - such as the Gene Ontology term 'hydrolase activity, acting on ester bonds' or the LMO term 'receptor-mediated extra-hepatic cellular uptake' - which do not appear literally in text seem to be a key point for the further improvement of the results.

The terms that were absent from the automatically generated terminologies were grouped into five categories of ranging difficulty: *rarely occurring terms* ('test person', 'experimentee'), *rarely occurring variants of terms* ('insuline resistant', 'slo syndrome'), *very long terms* ('receptor-mediated extra-hepatic cellular uptake', 'predominance of large low-density lipoprotein particles'), *combinations of terms/variants* ('elevated plasma-tg level') and, finally, *terms that should normally be easily found* ('type-II diabetic', 'diabetes type I', 'apolipoprotein-c').

Terms from the latter category were terms that appear often in PubMed and should normally be identified, but were probably absent from the document set used to automatically generate the terminology. The document selection seems to be another key point for producing terminologies that can cover a whole domain. Much attention needs to be put at the first step in order to collect documents that are specific enough and include detailed terminology, but also general enough in order to include basic terms of interest.

Remaining open problems contain the selection of suitable corpora for term recognition as well as generation of composite terms (such as GO term 'hydrolase activity, acting on ester bonds') from basic ones.

Appendix A

Word Sense Disambiguation collected corpora

The WSD collected corpora used in the experiments can be found under:

http://www.ncbi.nlm.nih.gov/pmc/articles/PMC2663782/bin/1471-2105-10-28-S1.txt

The three corpora (High quality/Low quantity corpus; Medium quality/Medium quantity corpus; Low quality/High quantity corpus) are given in the form of PubMed identifiers (PMID) for True/False cases for the 7 ambiguous terms examined (GO/MeSH/UMLS identifiers are also given).

Appendix B

User-centered Evaluation of Semantic Browsers

B.1 Methods

To prove or disprove the hypotheses H1-6, the following questions were considered:

Mobility and travel within the system

- **O1**: time taken for users to find information or perform tasks
- **O2**: pathway taken to find information or perform tasks
- **O3**: use of semantic links compared with non-semantic links

  a) Do users use semantic links?

  b) Which semantic links are they using - tree, semantic relationships, etc.?

  c) What percentages of links are non-semantic and semantic?

User attitude and satisfaction

- **O4**: user satisfaction with the ease of use of the system
- **O5**: user attitudes to the availability of semantic links and ranking
- **O6**: user understanding of the SWB:

  a) Does the user think it helps him/her find information or complete tasks?

  b) Does the user understand how to use the SWB to find such information or complete such tasks?

Sample populations

Table B.1 shows the target population and the control platforms and intervention SWBs each population used. The initial aim was to recruit 10 users per intervention SWB. Group A1 included mainly infectious disease clinicians and group A2 molecular biologists.

Population	A1 – infectious disease practitioners	A2 – molecular biologists
Control	NeLI	PubMed
Intervention	COHSE-NeLI COHSE-NeLI	GoPubMed/GoGene GoPubMed/GoGene

Tab. B.1: Sample populations for the evaluation.

Step	SC users starting with the control platform	SI users starting with the intervention SWB
1	Pre-questionnaire regarding user demographics and previous experience with the control platform	Web server log collection
2	Task carried out using control platform	Task carried out using intervention SWB
3	Post-task questionnaire	
4	Repeat steps 2 and 3 until half of the tasks are completed	
5	Task carried out using intervention SWB	Task carried out using control platform
6	Post-task questionnaire	
7	Repeat steps 5 and 6 until all of the tasks are completed	
8	Post-questionnaire regarding user satisfaction and attitude	
9	Semi-structured interviews (workshops only)	

Tab. B.2: Evaluation structure.

Settings

The evaluation was carried out both **online** and in **workshops**. The online evaluation was necessary to evaluate the SWBs in real-world conditions and to increase the number of participants. Because remote users' questionnaire answers may misrepresent their experience, their behaviour was tracked with Web server logs. The workshop evaluation was necessary to observe users' behaviour and collect further qualitative data with semi-structured interviews.

Structure

Although users would become more familiar with the SWB by doing more tasks – and potentially give more accurate feedback – time constraints were recognised as a possible problem. To manage the risk that online users would fail to complete a lengthy evaluation, a short format, with fewer tasks, was devised for the online evaluation, and the long format, with more tasks, was used in the workshops. A complete list of tasks is provided in Section B.3. Table B.2 shows the final structure of the evaluation.

Control/intervention split

Instead of splitting the users into a control group and an intervention group, the evaluation was structured so that each user would use both the control and the intervention systems. It was also decided that, for each respective SWB, all of the users would be given the same set of tasks to do in the same order. The split would be implemented through counterbalancing, with some users doing the first half of the tasks using the control platform and the second half of the tasks using the intervention SWBs, and other users vice versa.

Data collection

Data collection from three sources was planned. The first source was the Web server logs collected automatically as users navigated the website. The second source was the pre-evaluation, post-task, and post-evaluation questionnaires as described in Table B.2. The third source was the semi-structured interviews to be conducted at the workshops.

Comparison with other evaluations Hoeber and Yang (2007) have identified a number of choices faced by designers of user evaluations for Web search interfaces.

Number of interfaces evaluated by each user Because the goal was to compare the SWBs with non-semantic browsers rather than with each other, and because of anticipated time constraints on users, we chose a *within-subjects* rather than a *between-subjects* experiment design (exposing each user to both the control and intervention interfaces, rather than to just one interface). Because each intervention SWB was an enhancement to its control platform, a risk of bias typical of within- subjects experiments remained: users might apply knowledge of one interface to the next. This was handled by counterbalancing the order in which users were exposed to the control and intervention systems (see "Control/intervention split" earlier). For the same reason we decided to use multiple tasks, rather than repetition of the same task.

Task definition Another choice is between allowing users to choose their own search topics, or predefining tasks for them. We chose to predefine our own tasks because user-defined tasks would have made it difficult to define completion criteria. Sets of predefined tasks such as the TREC 2005 HARD Track[1] are available, but not necessarily applicable to the biomedical domain nor to the features of the SWBs.

Uniformity of result sets Ensuring that all of the SWBs provide access to the same result set (Hoeber and Yang, 2007) was not an issue for our study as the SWBs were being compared to non-semantic systems, rather than to each other. Whereas CORESE-NeLI retrieves results only from the NeLI DL, and GoPubMed/GoGene retrieves results only from PubMed, the purpose of COHSE is to provide external links, so result sets between control and intervention could not have been uniform for all SWBs.

Elicitation of relevance ratings Rather than require users to rate the relevance of individual documents or rank their top results (Nowicki, 2003; Su, 2003; Vaughan, 2004), we decided to use the post-questionnaire to capture subjective ratings of the overall relevance of results, and use the weblogs to record which documents were actually viewed.

Completion criteria For Group A1, it was decided to have a single target document for each task, considered completed by the user's visiting that document. Because links on PubMed change frequently, there could be no specific target documents for Group A2, so the GoPubMed/GoGene task completion criterion would be the user's subjective perception of having found the answer. Asking participants to print out the results (Su, 2003) would have been unfeasible, especially for COHSE's link boxes.

Time to completion The weblogs would capture objective measures, and the post-task and post-evaluation questionnaires subjective perceptions, of time to completion.

[1] http://trec.nist.gov/data/t14_hard.html

Capturing responses to questionnaires We chose Web rather than paper forms for the questionnaires, to accommodate remote users and to maintain participants' focus and facilitate data analysis. Verbal protocols were ruled out; measures of intuitiveness might also have been biased by users' overhearing each others' comments. It would have been unfeasible – and a distraction from the SWBs' functionality – to add features in the interface for capturing users' opinions (Hu *et al.*, 1999).

Implementation of data collection Data was collected from the three planned sources. The Web server logs provided answers for O1, O2, and O3. The questionnaires provided answers for O4, O5, and O6a) and the interviews provided answers for O6b).

Implementation of questionnaires All of the evaluations began with a pre-questionnaire for demographic information (occupation/main degree, length of professional experience, preferred online research sources, experience of the control platform). Each task was followed by a post-task questionnaire containing 2 questions: *How well did the information you found answer the question?* (answer choices: Not at all, Partially, Fully) and *Was finding the answer in the information returned by the search engine:* (answer choices: Hard, Neither Hard nor Easy, Easy). Each evaluation ended with a post-questionnaire about ease of use of the system, information findability, relevance of information returned, overall system speed, and overall system likeability. Except for one question relating only to the SWB, each question required 2 answers: one for the control platform, and one for the intervention SWB (Shneiderman, 1992). An example is: *I found the system unnecessarily complex (Brooke, 1996). a) Unmodified system (NeLI alone) b) Modified system (NeLI + [SWB]).* The answer choices were on Likert scales, commonly used in questionnaires to specify a level of agreement with a statement. An example would be a scale from 1 (strongly disagree) to 5 (strongly agree). Most of the answer choices for Group A1 were on a scale of 1 (worst) to 5 (best), and most of the answer choices for Group A2 were on a scale of 1 (worst) to 10 (best). Group A2 had additional questions about the functionality of GoPubMed/GoGene. The complete questionnaires are shown in Section B.4.

Implementation of semi-structured interviews Workshop participants were interviewed where possible, using a loose structure with introductory questions (name, job title, etc.) followed by questions about the user experience such as *"What would make you want to use [the SWB] regularly?"*, a question which was worded to overcome reluctance to give negative feedback by reframing it as suggestions for improvement. The interview structure as well as notes from the interview are shown in Section B.5.

Implementation of web server logs The server logs of the respondents' actions were analysed using a combination of logs produced by the SWBs and the server at City University, which hosted the online evaluation questionnaire and the NeLI website. Each respondent was assigned a unique identifier (uID) at the start of the evaluation, which was then passed between each page of the online evaluation and the SWB and the NeLI or GoPubMed website.

Implementation of tasks The COHSE and the CORESE-based SWB tasks were defined by one of the evaluators, a lay person with no medical knowledge, and reviewed by a colleague with medical expertise. The goal was for the framework to be applicable to any SWB. While we believe that this goal has potentially been met by the framework as a whole, one part of the evaluation process cannot be generalised: the task definition. Since each of the SWBs was different in nature, the same tasks would not have been appropriate for each. COHSE uses the NeLI vocabulary to present links *external* to the NeLI DL. The CORESE-based SWB presents a graph of the vocabulary for navigation *within* the NeLI DL and sorts the search results according to the hierarchical position in the vocabulary of the relevant terms found in the documents. GoPubMed and GoGene are search technologies applied to the PubMed search engine. For COHSE and the CORESE-based SWB, the tasks were counterbalanced with

users with even-numbered uIDs starting with the intervention SWB, and those with odd- numbered uIDs starting with the control platform. Thus, the same task was sometimes answered with the SWB and sometimes with NeLI. The answer to each task was always located in a single target document. A task was considered complete when the user felt that the answer had been found, whereupon a "Completed" button on the task page took them to the post-task questions. Users were asked not to spend more than 5 minutes on any one question. We were conscious of the contrivance inherent in posing questions the exact answer to which could only be found in a single target document. However, the need for authenticity had to be balanced against the need to know whether or not the test had been passed; detecting whether a single target document was found was the most unequivocal way to achieve that. The answers also needed to be detailed enough that participants would be unlikely to know every detail from memory, and so mark the question as answered without first searching for the answer. To counterbalance this contrivance, the questions needed to be general enough to be partially answerable through NeLI alone, and this had to be demonstrable in search results *without* a single specific target document. For COHSE, the target document was only reachable through a prominently visible link in a link box. The link box would appear when the user clicked on a specific related term highlighted by COHSE and found either on the NeLI home page or after searching for relevant terms in the NeLI website. An example is ***What kind of certificate should be used for documenting yellow fever vaccination? Have there been any changes to the format in the past two years?*** For the CORESE-based SWB, the target document either could not be found using a search of the NeLI website alone (at least, not by using predictable search terms), or else the target could be found via NeLI alone, but ranked lower than 20 in the results. An example is ***What are the recommended guidelines for hygienic cleaning of surfaces after flooding?*** Dedicated tasks were also devised for GoPubMed (online) and GoGene and an extended GoPubMed (workshop). These were not counterbalanced but were in the sequence described in the paragraph "Format". They were designed to avoid, as much as possible, bias in favour of the SWB. Target documents were not specified for the GoPubMed evaluation because the results change on PubMed so frequently. The participants were told not to spend more than 8 minutes to answer each task. An example is ***What is the main role of the gene MMS2? Name 3 genes related to it in literature (PubMed)***; and ***What other genes are related to Shh in literature? Name 3 of them (GoGene)***. For the workshop, another PubMed task was ***Can you find any conserved domain information on Rab5?*** and an extended GoPubMed task was ***Can you find any conserved domain information on Apc11?***

B.2 Results

Recruitment of online participants The online evaluation ran from December 2008 to March 2009. Users were recruited through advertisements and newsletter bulletins circulated to the mailing lists of NeLI and its companion site NRIC (http://www.nric.org.uk, the National Resource for Infection Control), and news bulletins on the sites' home pages.

Recruitment of workshop participants (Group A1) The workshops for COHSE-NeLI and CORESE-NeLI (Group A1) took place in London, all at City University, except for one which was hosted by the Health Protection Agency (HPA) Centre for Infections. Recruitment was through invitations circulated to the NeLI and NRIC mailing lists, to the HPA, to the Infection Prevention Society, and to other organisations through the evaluators' professional contacts. The original plan was to hold one 2-hour workshop at a fixed date and time at City University followed by one hour for lunch and semi-structured interviews. However, due to the constraints on clinicians' time, acceptances were few and cancellations many. Because of these difficulties, a workshop was planned at the HPA Centre for Infections, where workstations were reserved in the library for staff to participate throughout the day. The event was advertised a week in advance, using posters and internal news systems; fliers and a stand were used in the canteen on the evaluation day. In this way, 14 participants were recruited, prompting the use of similar strategies at the two subsequent workshops that were held at City University.

Recruitment of workshop participants (Group A2) One workshop was held for GoGene and the extended GoPubMed (Group A2), at the Biotechnology Centre of the Technische Universität in Dresden, where postgraduate students constituted a source of real-world users. A successful recruitment strategy was through personal contacts of one of the evaluators, admittedly introducing some risk of bias, but securing attendance of a higher number of real-world users.

Demographics The following section describes the results for each SWB. Table B.3 shows the number of participants from each target and non-target audience. Groups with a majority of participants from the target audience were the COHSE-NeLI online group, the CORESE-NeLI workshop group, and the GoGene/extended GoPubMed workshop group. Only 2 of the CORESE-NeLI online group completed any tasks, and one of those dropped out after the control tasks, leaving the intervention tasks untouched. A possible explanation is that CORESE requires installation of a plugin, which may have been off-putting to this user group.

Format All online evaluations were held in the **short format** of 4 tasks. For Group A1, the **long format** of 10 tasks for workshops proved too time-consuming and was abandoned in favour of the **short format**. For Group A2, the **long format** was used as planned, with 11 tasks instead of 10; 2 hours were allowed and proved sufficient. The tasks for Group A1 were counterbalanced as planned. The tasks for Group A2 were not counterbalanced: some tasks were answered with control only and some with the intervention SWB only.

Objectives

O1: time taken for users to find information or perform tasks The time taken per task was calculated from the online evaluation logs using the difference between task page and question page loading times. In the process it was noted that some users had not completed all the tasks, and others had completed all the tasks, but within an unrealistic timescale (e.g. more than 2 tasks completed under 60 seconds). These users were removed from the log evaluation. Additionally, logging was unavailable for the extended GoPubMed, so O1, O2, and O3 could not be answered for this SWB and log analysis

SWB	Setting	# participants	Occupation	# participants
COHSE-NeLI	**Online**	***39***	Medical	21
			Scientific	6
			Other	12
	Workshop	***28***	Medical	4
			Information	10
			Student	14
CORESE-NeLI	**Online**	***4***	Medical	3
			Researcher	1
	Workshop	***14***	Medical Biological	2 6
			Information	3
			Unspecified	3
eliminated for completing tasks unrealistically quickly				2
GoPubMed	**Online**	***141***	Biology	21
			Chemistry	1
			Physics	2
			Other	113
GoGene / extended GoPubMed	**Workshop**	***14***	Other	4 (of whom 3 scientists)
			Biology	8

Tab. B.3: User demographics.

GoPubMed	GoGene	COHSE	CORESE	PubMed	NeLI
126	229	478	266	194	387

Tab. B.4: Average time for all tasks on each system in seconds.

of the extended GoPubMed is not included. Table B.4 shows the average times spent using each system and the PubMed and NeLI websites. This suggests that GoPubMed tasks were the quickest in just over 2 minutes. The slowest tasks were for COHSE in just under 8 minutes.

O2: pathway taken to find information or perform tasks For the COHSE evaluation, none of the 28 users included in the log analysis for the short format found the target documents via COHSE. For the CORESE evaluation, 11 users were included in the log analysis, of whom 8 started with CORESE and 3 with NeLI. Table B.5 shows the number of users who viewed the target documents via the CORESE-based SWB. This shows that there were very few users who actually found the target documents with the CORESE- based SWB. As stated, for the GoPubMed/GoGene there were no specific target documents for these SWBs. Logs of users' actions were however recorded to show how much a user was interacting with the site during the tasks. Users performed up to 40 actions whilst looking for the information and the majority of respondents used less than 15 actions to find the information on GoPubMed and less than 25 on GoGene. Access to the server logs for PubMed was not available for this evaluation.

O3: use of semantic links compared with non-semantic links For COHSE, an indication of the use of semantic links is the number of times a highlighted term is clicked and the link box activated. A further indication is the number of views of external sites via COHSE. 6 users did not click on any highlighted terms and therefore did not use any of the semantic features. In the short format, 132 sites external to NeLI were viewed from 97 link box activations. The largest number of link box activations

Task 1	Task 2	Task 3	Task 4
2/8	2/8	1/3	2/3

Tab. B.5: Proportion of users who viewed the target document for each task.

per user was 15, the lowest 1; the median was 4 and the mode, 3.

Of those users who viewed external pages via COHSE-NeLI, the largest number of views per user was 42 for the short format and 192 for the long. The lowest for the short format was 1; for the long 24; the median was 5.5 for the short format, 82.5 for the long. The mode for the short format was 1; the long format had no mode. For CORESE-NeLI it was not possible to directly compare the use of semantic links with non-semantic links because all of the links that a user interacts with on the CORESE-based SWB can be classed as semantic. There were however 325 searches via the CORESE-based SWB compared to 91 searches via NeLI, suggesting that users interacted with the CORESE-based SWB more than they would a standard website. For GoPubMed, around 46% of users used the semantic features at least once, but as an overall percentage of activity, semantic activity was relatively low. For GoGene, from a total of 270 recorded actions, 73 were classed as semantic actions (27%), generated by 10 individual users (one user was not found in the logs).

O4: user satisfaction with the ease of use of the system

Usability COHSE scored 1 point higher (on a scale of 1=worst to 5=best) than control for complexity, the CORESE-based SWB 1 point lower, GoPubMed/GoGene the same. COHSE also scored as 2 points (out of 5) more satisfying than the control platform. The CORESE-based SWB scored worse than control for rigidity. GoPubMed/GoGene scored 3 points higher (on a scale of 1=hardest to 10=easiest) than control for ease of use; there was no difference for COHSE and the CORESE- based SWB. GoPubMed scored 1 point higher (out of 10) than control for provision of help, with no equivalent question for the other SWBs.

Overall likeability of the system COHSE scored better than control in 1 of the 3 questions posed (I think that I would like to use this system frequently), while the CORESE-based SWB scored worse than control for the same question, and GoPubMed/GoGene scored 3 points higher than PubMed (out of 10), a greater difference than the equivalent superior score for COHSE.

Overall system speed COHSE scored worse than control for speed. GoPubMed and GoGene scored the same as control.

GoPubMed/GoGene functionality Though there are no equivalent questions for the other SWBs, the functionality of GoPubMed and GoGene was well regarded.

O5: user attitudes to the availability of semantic links and ranking COHSE and the CORESE-based SWB both scored better than control for absence of irrelevant results. The CORESE-based SWB scored better than control for relevance of results, while COHSE scored the same. GoPubMed/GoGene also scored better (by 3 points on a scale of 1=worst to 10=best) in the equivalent measures to those in which COHSE and the CORESE-based SWB triumphed. While COHSE scored best for absence of irrelevant results, GoPubMed/GoGene scored better than the CORESE- based SWB in this respect. GoPubMed/GoGene had the best scores for relevance of results.

O6: user understanding of the SWB

a) Does the user think it helps him/her find information or complete tasks? The CORESE-based SWB scored 3 points worse (out of 5) than control for ease of finding answers (Table B.7), for which GoPubMed/GoGene scored 2 points better (out of 10) than control, and 3 points better for speed of finding answers (Table B.8).

	Mode
Did you find the highlighting of ontology terms helpful? (Yes/No)	Yes
Did you get an overview over your search results from the tree on the left? (Yes/No)	Yes
Did you manage to navigate efficiently through the tree? (Yes/No)	Yes
Did you find any papers you would probably have missed with PubMed? (Yes/No)	Yes

Tab. B.6: Mode Scores for GoPubMed/GoGene functionality (Yes/No).

	COHSE	CORESE
Speed of finding answers in info returned: Was COHSE or CORESE rated higher or lower than NeLI?	0	0
Ease of finding answers in info returned: Was COHSE or CORESE rated higher or lower than NeLI?	0	-3

Tab. B.7: Findability of COHSE and the CORESE-based SWB: mode differences (scale 1 bad - 5 good).

b) Does the user understand how to use the SWB to find such information or complete such tasks? To test intuitiveness, all of the online evaluations, and the early Group A1 workshops, opened with minimal introduction. It quickly emerged that many users could not tell the control and intervention systems apart, giving detailed feedback on NeLI while assuming that the COHSE link boxes were advertisements or error messages. Consequently, introductory presentations were shown to each user at subsequent workshops. This reduced confusion, but users still said more introduction was needed. Even users who could tell NeLI apart from the SWBs complained of distraction by the NeLI website's user interface. A widely familiar control platform such as Google would have increased the contrast and foregrounded the benefits of COHSE in particular. Users did not grasp the nature of the CORESE-based SWB at all, assuming it to be a keyword search with a graph attached. A detailed introduction would probably have greatly improved users' opinions. The GoPubMed workshop opened with a 20-minute introductory lecture, and PubMed is a widely familiar system to use as a control. The difficulties encountered by Group A2 were noted as being generally more trivial than those found by Group A1.

Overall post-questionnaire scores In no case did GoPubMed/GoGene receive worse mode scores than control, whereas COHSE and the CORESE-based SWB received several lesser modal scores.

	Mode
Speed of finding answers in info returned: Was GoPubMed/GoGene rated higher or lower than PubMed?	2
Ease of finding answers in info returned: Was GoPubMed/GoGene rated higher or lower than PubMed?	3

Tab. B.8: Findability of GoPubMed/GoGene: mode differences (scale 1 bad - 5 good).

B.3 List of Tasks

COHSE-NeLI evaluation tasks

Short format (4 tasks)

1. What percentage of viral encephalitis and meningitis cases in the UK are of undetermined aetiology?
2. Were there any travel restrictions to Turkey at the time when it was affected by avian influenza? In what areas of Turkey did Europe's first confirmed human cases of avian influenza occur?
3. What is an example of a chemical that can cause or exacerbate acne?
4. What kind of certificate should be used for documenting yellow fever vaccination? Have there been any changes to the format in the past two years?

Long Format (10 tasks)

The long COHSE evaluation consisted of the above four questions, plus the following six.

1. Has the risk of dengue virus in Cambodia increased or decreased over the last two years?
2. Are there any infections that pose more of a threat in Latvia than they do in the UK? If so, one would not expect malaria to be among them, but could this expectation be wrong?
3. What are three chemicals that have successfully been used for decontamination of areas affected by anthrax?
4. How much risk of skin damage can the Mediterranean sun present to a typical redhead?
5. Which particular diseases are the focus of infection prevention efforts in prisons?
6. If a patient presented with uncomplicated dyspepsia that did not respond to lifestyle changes, what commonly used non-invasive test would be the most likely to give you a false positive for helicobacter pylori?

CORESE-Based SWB evaluation tasks

1. What are three disease prevention efforts in the USA to prevent and control the spread of viral haemorrhagic fever through rodent vectors?
2. In the Netherlands in 2007, there was a cluster of infections in a care home caused by poor infection control practices. What was the infection and what (specifically) was its most likely cause?
3. What are the recommended guidelines for hygienic cleaning of surfaces after flooding?
4. Which are the four main types of hospital unit within which outbreaks of VRE have been reported?

GoPubMed/GoGene evaluation tasks

Short format (4 tasks, PubMed vs GoPubMed)

Answered using PubMed

1. Which particular diseases are associated most often with HIV?
2. What is the role of MMS2 in cancer?

Answered using GoPubMed

1. How does BARD1 regulate BRCA1 activity?
2. How rare/prevalent is Paget's disease?

Long format (11 tasks, PubMed vs GoGene and PubMed vs extended GoPubMed)

Answered using PubMed

1. What is the main role of the gene MMS2? Name 3 genes related to it in literature.
2. How are MMS2 and UBC13 related to each other?

Answered using GoGene

1. Name 4 organisms that have an MMS2 gene.
2. How do mutations in "Sonic Hedgehog" genes affect developmental disorders?
3. What other genes are related to Shh in literature? Name 3 of them.
4. Are there any GO terms manually assigned to Shh? Name 3 of them.

Answered using PubMed

1. Which biological process is Rab5 involved in?
2. Can you find any conserved domain information on Rab5?

Answered using the extended GoPubMed

1. Which biological process is Apc11 involved in?
2. Can you find any conserved domain information on Apc11?
3. Which pathway are the zebrafish genes Her1, Her7 and DeltaC involved in?

B.4 Questionnaires

All the pre-questionnaires, post-task questionnaires, and post-questionnaires for both Group A1 and Group A2 can be found under:

http://www.ncbi.nlm.nih.gov/pmc/articles/PMC2755822/bin/1471-2105-10-S10-S14-S2.txt

B.5 Semi-structured interviews

The questions given to all the interviewers to guide the semi-structured interviews can be found under:

http://www.ncbi.nlm.nih.gov/pmc/articles/PMC2755822/bin/1471-2105-10-S10-S14-S3.txt

Notes from semi-structured interviews, Group A1

Notes from all the semi-structured interviews for Group A1 can be found under:

http://www.ncbi.nlm.nih.gov/pmc/articles/PMC2755822/bin/1471-2105-10-S10-S14-S4.txt

Notes from semi-structured interviews, Group A2

Notes from all the semi-structured interviews for Group A2 can be found under:

http://www.ncbi.nlm.nih.gov/pmc/articles/PMC2755822/bin/1471-2105-10-S10-S14-S5.txt

References

Agirre, E. and Martinez, D. (2000). Exploring automatic word sense disambiguation with decision lists and the web. In *Proceedings of the 18th International Conference on Computational Linguistics (COLING, Saarbrücken, Germany)*, pages 11–19.

Agirre, E. and Stevenson, M. (2006). *Word Sense Disambiguation: Algorithms and Applications*, chapter Knowledge Sources for WSD, pages 217–251. Springer.

Alexopoulou, D., Andreopoulos, B., Hakenberg, J., and Schroeder, M. (2007a). Word-sense disambiguation in biomedical ontologies with term co-occurrence analysis. In *German Conference on Bioinformatics 2007 (GCB07), Potsdam, Germany*. Poster presentation.

Alexopoulou, D., Andreopoulos, B., Hakenberg, J., and Schroeder, M. (2007b). Word-sense disambiguation in biomedical ontologies with term co-occurrence analysis. In *Physical and Chemical Foundations of Bioinformatics Methods International workshop, Dresden, Germany*. Poster presentation.

Alexopoulou, D., Andreopoulos, B., Dietze, H., Doms, A., Gandon, F., Hakenberg, J., Khelif, K., Schroeder, M., and Wächter, T. (2008a). Biomedical word sense disambiguation with ontologies and metadata. In *3rd conference of the Hellenic Society for Computational Biology and Bioinformatics*. Short Oral communication.

Alexopoulou, D., Wächter, T., Pickersgill, L., Eyre, C., and Schroeder, M. (2008b). Terminologies for text-mining; an experiment in the lipoprotein metabolism domain. *BMC Bioinformatics*, **9 Suppl 4**, S2.

Alexopoulou, D., Andreopoulos, B., Dietze, H., Doms, A., Gandon, F., Hakenberg, J., Khelif, K., Schroeder, M., and Wächter, T. (2009). Biomedical word sense disambiguation with ontologies and meta-data: automation meets accuracy. *BMC Bioinformatics*, **10**(1), 28.

ALPAC (1966). Language and machine: Computers in translation and linguistics. A report by the Automatic Language Processing Advisory Committee, Division of Behavioral Sciences, National Research Council. Washington, D.C.: National Academy of Sciences.

Altun, Z. and Hall, D., editors (2002-2006). *WormAtlas*.

Andreopoulos, B., An, A., and Wang, X. (2007). Hierarchical density-based clustering of categorical data and a simplification. In *PAKDD*, pages 11–22.

Andreopoulos, B., Alexopoulou, D., and Schroeder, M. (2008). Word sense disambiguation in biomedical ontologies with term co-occurrence analysis and document clustering. *International Journal of Data Mining and Bioinformatics*, **2**(3), 193–215. (Special Issue on Text Mining and Information Retrieval).

Andritsos, P., Tsaparas, P., and abd K Sevcik, R. M. (2004). Limbo: Scalable clustering of categorical data. In *Ninth International Conference on Extending Database Technology (EDBT)*, pages 123–146, Heraklion, Greece.

Aranguren, M. E., Bechhofer, S., Lord, P., Sattler, U., and Stevens, R. (2007). Understanding and using the meaning of statements in a bio-ontology: recasting the gene ontology in owl. *BMC Bioinformatics*, **8**, 57.

Arikuma, T., Yoshikawa, S., Azuma, R., Watanabe, K., Matsumura, K., and Konagaya, A. (2008). Drug interaction prediction using ontology-driven hypothetical assertion framework for pathway generation followed by numerical simulation. *BMC Bioinformatics*, **9**(Suppl 6), S11.

Ashburner, M., Ball, C., Blake, J., Botstein, D., Butler, H., Cherry, J., Davis, A., Dolinski, K., Dwight, S., Eppig, J., Harris, M., Hill, D., Issel-Tarver, L., Kasarskis, A., Lewis, S., Matese, J., Richardson, J., Ringwald, M., Rubin, G., and Sherlock, G. (2000). Gene ontology: tool for the unification of biology. the gene ontology consortium. *Nat Genet*, **25**(1), 25–9.

Azuaje, F., Wang, H., and Bodenreider, O. (2005). Ontology-driven similarity approaches to supporting gene functional assessment. In *Proceedings of the ISMB'2005 SIG meeting on Bio-ontologies*, pages 9–10.

Azuma, N., Tadokoro, K., Asaka, A., Yamada, M., Yamaguchi, Y., Handa, H., Matsushima, S., Watanabe, T., Kida, Y., Ogura, T., Torii, M., Shimamura, K., and Nakafuku, M. (2005). Transdifferentiation of the retinal pigment epithelia to the neural retina by transfer of the pax6 transcriptional factor. *Hum Mol Genet*, **14**(8), 1059–68.

Baker, P., Goble, C., Bechhofer, S., Paton, N., Stevens, R., and Brass, A. (1999). An ontology for bioinformatics applications. *Bioinformatics*, **15**(6), 510–20.

Baldock, R. A., Bard, J. B. L., Burger, A., Burton, N., Christiansen, J., Feng, G., Hill, B., Houghton, D., Kaufman, M., Rao, J., Sharpe, J., Ross, A., Stevenson, P., Venkataraman, S., Waterhouse, A., Yang, Y., and Davidson, D. R. (2003). Emap and emage: a framework for understanding spatially organized data. *Neuroinformatics*, **1**(4), 309–25.

Banerjee, S. and Pedersen, T. (2003). Extended gloss overlaps as a measure of semantic relatedness. In *Proceedings of the 18th International Joint Conference on Artificial Intelligence (IJCAI, Acapulco, Mexico)*, pages 805–810.

Banville, D. L. (2009). Mining chemical and biological information from the drug literature. *Curr Opin Drug Discov Devel*, **12**(3), 376–87.

Bard, J., Rhee, S., and Ashburner, M. (2005). An ontology for cell types. *Genome Biol*, **6**(2), R21.

Bard, J. L., Kaufman, M. H., Dubreuil, C., Brune, R. M., Burger, A., Baldock, R. A., and Davidson, D. R. (1998). An internet-accessible database of mouse developmental anatomy based on a systematic nomenclature. *Mech Dev*, **74**(1-2), 111–20.

Berger, A. L., Pietra, V. J. D., and Pietra, S. A. D. (1996). A maximum entropy approach to natural language processing. *Comput. Linguist.*, **22**(1), 39–71.

Berneis, K. and Rizzo, M. (2005). Ldl size: does it matter? *Swiss Med Wkly*, **134**(49-50), 720–4.

Berners-Lee, T., Hendler, J., and Lassila, O. (2001). The semantic web. *Scientific American*, **284**(5), 34–43.

Berners-Lee, T., Hollenbach, J., Lu, K., Presbrey, J., Prud'ommeaux, E., and Schraefel, M. (2007). Tabulator redux: Writing into the semantic web. Technical Report ECSIAMeprint14773, Electronics and Computer Science, University of Southampton.

Blake, J. A. and Bult, C. J. (2006). Beyond the data deluge: data integration and bio-ontologies. *J Biomed Inform*, **39**(3), 314–320.

Bodenreider, O. (2004). The unified medical language system (UMLS): integrating biomedical terminology. *Nucleic Acids Research*, **32**(D267-70).

Bodenreider, O. and Stevens, R. (2006). Bio-ontologies: current trends and future directions. *Brief Bioinform*, **7**(3), 256–274.

Boser, B., Guyon, I., and Vapnik, V. (1992). A training algorithm for optimal margin classifiers. In *Proceedings of the 5th Annual Workshop on Computational Learning Theory (Pittsburgh, PA)*, pages 144–152.

Boyack, K. W. (2004). Mapping knowledge domains: characterizing pnas. *Proc Natl Acad Sci U S A*, **101 Suppl 1**, 5192–9.

Brants, T. and Franz, A. (2006). Web 1t 5-gram, ver. 1. Technical report, Linguistic Data Consortium, Philadelphia, PA.

Brin, S. and Page, M. (1998). Anatomy of a large-scale hypertextual web search engine. In *Proceedings of the 7th Conference on World Wide Web (Brisbane, Australia)*, pages 107–117.

Brinckmann, A., Rüther, K., Williamson, K., Lorenz, B., Lucke, B., Nürnberg, P., Trijbels, F., Janssen, A., and Schuelke, M. (2007). De novo double mutation in pax6 and mtdna trna(lys) associated with atypical aniridia and mitochondrial disease. *J Mol Med*, **85**(2), 163–8.

Brooke, J. (1996). *Usability Evaluation in Industry*, chapter SUS - A quick and dirty usability scale, pages 189–194. Taylor and Francis.

Bruce, R. and Wiebe, J. (1994). Word-sense disambiguation using decomposable models. In *Proceedings of the 32nd Annual Meeting of the Association for Computational Linguistics*, pages 139–145, ACL, Las Cruces, NM.

Bruce, R. and Wiebe, J. (1999). Decomposable modeling in natural language processing. *Comput. Ling.*, **25**(2), 195–207.

Budanitsky, A. and Hirst, G. (2006). Evaluating WordNet-based Measures of Lexical Semantic Relatedness. *Computational Linguistics*, **32**(1), 13–47.

Buscaldi, D., Rosso, P., Pla, F., Segarra, E., and Sanchis, E. (2006). Verb sense disambiguation using support vector machines: impact of wordnet-extracted features. In *Proc. Int. Conf. on Comput. Linguistics and Intelligent Text Processing, CICLing-2006*, pages 192–195. Springer Verlag.

Camon, E., Magrane, M., Barrell, D., Lee, V., Dimmer, E., Maslen, J., Binns, D., Harte, N., Lopez, R., and Apweiler, R. (2004). The Gene Ontology Annotation (GOA) Database: sharing knowledge in Uniprot with Gene Ontology. *Nucleic Acids Res*, **32**(Database issue), D262–D266.

Camon, E., Barrell, D., Dimmer, E., Lee, V., Magrane, M., Maslen, J., Binns, D., and Apweiler, R. (2005). An evaluation of GO annotation retrieval for BioCreAtIvE and GOA. *BMC Bioinformatics*, **6**(Suppl 1), S17.

Camous, F., Blott, S., and Smeaton, A. F. (2007). Ontology-based medline document classification. In S. Hochreiter and R. Wagner, editors, *BIRD*, volume 4414 of *Lecture Notes in Computer Science*, pages 439–452. Springer.

Castro, A., Rocca-Serra, P., Stevens, R., Taylor, C., Nashar, K., Ragan, M., and Sansone, S. (2006). The use of concept maps during knowledge elicitation in ontology development processes–the nutrigenomics use case. *BMC Bioinformatics*, **7**, 267.

Chapman, R. (1977). *Roget's International Thesaurus*. New York: Harper and Row, 4th edition.

Charniak, E., Blaheta, D., Ge, N., Hall, K., Hale, J., and Johnson, M. (2000). Bllip 1987-89 WSJ corpus release 1. Technical report, Linguistic Data Consortium, Philadelphia, PA.

Cimiano, P. and Völker, J. (2005). Text2Onto - A Framework for Ontology Learning and Data-driven Change Discovery. In A. Montoyo, R. Muñoz, and E. Métais, editors, *NLDB*, volume 3513 of *Lecture Notes in Computer Science*, pages 227–238. Springer.

Clear, J. (1993). *The British national corpus*, pages 163–187. MIT Press On Technical Communication And Information Systems Series, Cambridge, MA, USA.

Consortium, T. F. (1998). Flybase – a drosophila database. *Nucl. Acids Res.*, **26**, 85–88.

Cottrell, G. (1989). *A Connectionist Approach to Word Sense Disambiguation*. Pitman, London, U.K.

Couto, F., Silva, M., and Coutinho, P. (2005). Finding genomic ontology terms in text using evidence content. *BMC Bioinformatics*, **6**(1), S21.

Daelemans, W., Bosch, A. V. D., and Zavrel, J. (1999). Forgetting exceptions is harmful in language learning. *Mach. Learn.*, **34**(1), 11–41.

Decadt, B., Hoste, V., Daelemans, W., and Bosch, A. V. D. (2004). GAMBL, genetic algorithm optimization of memory-based WSD. In *Proceedings of the 3rd International Workshop on the Evaluation of Systems for the Semantic Analysis of Text (Senseval-3, Barcelona, Spain)*, pages 108–112.

Degtyarenko, K., de Matos, P., Ennis, M., Hastings, J., Zbinden, M., McNaught, A., Alcntara, R., Darsow, M., Guedj, M., and Ashburner, M. (2008). Chebi: a database and ontology for chemical entities of biological interest. *Nucleic Acids Res*, **36**(Database issue), D344–50.

del Pozo, A., Pazos, F., and Valencia, A. (2008). Defining functional distances over Gene Ontology. *BMC Bioinformatics*, **9**(50).

DeLuca, D. S., Beisswanger, E., Wermter, J., Horn, P. A., Hahn, U., and Blasczyk, R. (2009). MaHCO: an ontology of the major histocompatibility complex for immunoinformatic applications and text mining. *Bioinformatics*, **25**(16), 2064–70.

Diallo, G., Kostkova, P., Jawaheer, G., Jupp, S., and Stevens, R. (2008a). Process of building a vocabulary for the infection domain. In *Proceedings of the 21st IEEE International Symposium on Computer-Based Medical Systems: 17-19 June 2008; Jyväskylä, Finland.*, pages 308–313. IEEE CS Press.

Diallo, G., Khelif, K., Corby, O., Kostkova, P., and Madle, G. (2008b). Semantic browsing of a domain specific resources: The corese-neli framework. In O. Hoeber and Y. Yao, editors, *Proceedings of the 2008 International Workshop on Web Information Retrieval Support Systems WIRSS 2008*, pages 50–54. IEEE. held in collaboration with the IEEE/WIC/ACM International Conference on Web Intelligence WI'08.

Dietze, H., Alexopoulou, D., Alvers, M. R., Barrio-Alvers, L., Andreopoulos, B., Doms, A., Hakenberg, J., Mönnich, J., Plake, C., Reischuck, A., Royer, L., Wächter, T., Zschunke, M., and Schroeder, M. (2008). *Bioinformatics for Systems Biology*, chapter GoPubMed: Exploring PubMed with Ontological Background Knowledge. The Human Press.

Doms, A. and Schroeder, M. (2005). Gopubmed: exploring pubmed with the gene ontology. *Nucl. Acids Res.*, **33**, W783–786.

Dorow, B. and Widdows, D. (2003). Discovering corpus-specific word senses. In *EACL '03: Proceedings of the tenth conference on European chapter of the Association for Computational Linguistics*, pages 79–82, Morristown, NJ, USA. Association for Computational Linguistics.

Edmonds, P. (2005). *The Elsevier Encyclopedia of Language and Linguistics*, chapter Lexical disambiguation, pages 607–23. Oxford: Elsevier, 2nd edition.

Edmonds, P. and Agirre, E. (2006). *Word Sense Disambiguation: Algorithms And Applications*. Springer Verlag.

Ehrler, F., Geissbuehler, A., Jimeno, A., and Ruch, P. (2005). Data-poor categorization and passage retrieval for gene ontology annotation in swiss-prot. *BMC Bioinformatics*, **6**(1), S23.

Eilbeck, K., Lewis, S. E., Mungall, C. J., Yandell, M., Stein, L., Durbin, R., and Ashburner, M. (2005). The sequence ontology: a tool for the unification of genome annotations. *Genome Biol*, **6**(5), R44.

Escudero, G., Màrquez, L., and Rigau, G. (2000a). Naive bayes and exemplar-based approaches to word sense disambiguation revisited. In *Proceedings of the 14th European Conference on Artificial Intelligence (ECAI, Berlin, Germany)*, pages 421–425.

Escudero, G., Màrquez, L., and Rigau, G. (2000b). On the portability and tuning of supervised word sense disambiguation. In *Proceedings of the Joint SIGDAT Conference on Empirical Methods in Natural Language Processing and Very Large Corpora (EMNLP/VLC, Hong Kong, China)*, pages 172–180.

Evsikov, A., de, V. W., Peaston, A., Radford, E., Fancher, K., Chen, F., Blake, J., Bult, C., Latham, K., Solter, D., and Knowles, B. (2004). Systems biology of the 2-cell mouse embryo. *Cytogenet Genome Res*, **105**(2-4), 240–50.

Farkas, R. (2008). The strength of co-authorship in gene name disambiguation. *BMC Bioinformatics*, **9**(69).

Fellbaum, C. (1998). *WordNet An Electronic Lexical Database*. MIT Press, USA.

Fensel, D., Horrocks, I., van Harmelen, F., McGuinness, D. L., and Patel-Schneider, P. F. (2001). Oil: An ontology infrastructure for the semantic web. *IEEE Intelligent Systems*, **16**(2), 38–45.

Frantzi, K., Ananiadou, S., and Mima, H. (2000). Automatic recognition of multi-word terms:. the c-value/nc-value method. *International Journal on Digital Libraries*, **V3**(2), 115–130.

Freund, Y. and Schapire, R. (1999). A short introduction to boosting. *J. Japanese Soci. Artific. Intell.*, **14**, 771–780.

Fujii, A., Inui, K., Tokunaga, T., and Tanaka, H. (1998). Selective sampling for example-based word sense disambiguation. *Computat. Ling.*, **24**(4), 573–598.

Gale, W., Church, K., and Yarowsky, D. (1992a). A method for disambiguating word senses in a corpus. *Comput. Human.*, **26**, 415–439.

Gale, W. A., Church, K. W., and Yarowsky, D. (1992b). One sense per discourse. In *HLT '91: Proceedings of the workshop on Speech and Natural Language*, pages 233–237, Morristown, NJ, USA. Association for Computational Linguistics.

Garfield, E. and Melino, G. (1997). The growth of the cell death field: an analysis from the isi-science citation index. *Cell Death Differ*, **4**(5), 352–61.

Gaudan, S., Kirsch, H., and Rebholz-Schuhmann, D. (2005). Resolving abbreviations to their senses in Medline. *Bioinformatics*, **21**(18), 3658–3664.

Ginter, F., Boberg, J., Jrvinen, J., and Salakoski, T. (2004). New techniques for disambiguation in natural language and their application to biological text. *J. Mach. Learn. Res.*, **5**, 605–621.

Gray, J. M. and de Lusignan, S. (1999). National electronic library for health (nelh). *BMJ*, **319**(7223), 1476–9.

Gruber, T. R. (1993). A translation approach to portable ontologies. *Knowledge Acquisition*, **5**(2), 199–220.

Grumbling, G. and Strelets, V. (2006). Flybase: anatomical data, images and queries. *Nucleic Acids Res*, **34**(Database issue), D484–8.

Guthrie, J. A., Guthrie, L., Wilks, Y., and Aidinejad, H. (1991). Subject dependent co-occurrence and word sense disambiguation. In *Proceedings of the 29th Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 146–152, Berkeley, California.

Hakenberg, J., Royer, L., Plake, C., Strobelt, H., and Schroeder, M. (2007). Me and my friends: gene mention normalization with background knowledge. In *Proceedings of the 2nd BioCreAtIvE Challenge Evaluation Workshop*, pages 141–144.

Hakenberg, J., Plake, C., Royer, L., Strobelt, H., Leser, U., and Schroeder, M. (2008). Gene mention normalization and interaction extraction with context models and sentence motifs. *Genome Biology*, **9**(Suppl 2), S14.

Halliday, M. and Hasan, R. (1976). *Cohesion in English*. Longman.

Hatzivassiloglou, V., Duboue, P. A., and Rzhetsky, A. (2001). Disambiguating proteins, genes, and RNA in text: a machine learning approach. *Bioinformatics*, **17**(suppl-1), S97–106.

Hearst, M. (1991). Noun homograph disambiguation using local context in large text corpora. In *Proceedings of the 7th Annual Conference of the UW Centre for the New OED and Text Research: Using Corpora (Oxford, U.K)*, pages 1–19.

Hearst, M. (1992). Automatic acquisition of hyponyms from large text corpora. In *Proceedings of the Fourteenth International Conference on Computational Linguistics: 23-28 August 1992; Nantes, France*, pages 539–545. Association for Computational Linguistics.

Henikoff, S. and Henikoff, J. (1992). Amino acid substitution matrices from protein blocks. *PNAS*, **89**(22), 10915–9.

Hirst, G. (1987). *Semantic Interpretation and the Resolution of Ambiguity*. Cambridge, UK: Cambridge University Press.

Hirst, G. and St-Onge, D. (1998). *WordNet: An Electronic Lexical Database*, chapter Lexical chains as representations of context for the detection and correction of malapropisms, pages 305–332. The MIT Press, Cambridge, MA.

Hoeber, O. and Yang, X. (2007). User-oriented evaluation methods for interactive web search interfaces. In Y. Li and V. Raghavan, editors, *2007 IEEE/WIC/ACM International Conferences on Web Intelligence and Intelligent Agent Technology Workshops (WI-IAT Workshops 2007): 2-5 November 2007; Silicon Valley, California, USA.*, pages 239–243. IEEE CS Press.

Hoehndorf, R., Loebe, F., Kelso, J., and Herre, H. (2007). Representing default knowledge in biomedical ontologies: application to the integration of anatomy and phenotype ontologies. *BMC Bioinformatics*, **8**, 377.

Hoste, V., Hendrickx, I., Daelemans, W., and Bosch, A. V. D. (2002). Parameter optimization for machine learning of word sense disambiguation. *J. Nat. Lang. Eng.*, **8**(4), 311–325.

Hu, P., Ma, P., and Chau, P. (1999). Evaluation of user interface designs for information retrieval systems: a computer-based experiment. *Decision Support Systems*, **27**, 125–143.

Huang, Z. (1998). Extensions to the k-means algorithm for clustering large data sets with categorical values. *Data Mining and Knowledge Discovery*, **2**(3), 283–304.

Humphrey, S. M., Rogers, W. J., Kilicoglu, H., Demner-Fushman, D., and Rindflesch, T. C. (2006). Word sense disambiguation by selecting the best semantic type based on Journal Descriptor Indexing: Preliminary experiment. *Journal of the American Society for Information Science and Technology*, **57**(1), 96–113.

Ide, N. and Véronis, J. (1998). Introduction to the special issue on word sense disambiguation: the state of the art. *Comput. Linguist.*, **24**(1), 2–40.

Jaiswal, P., Avraham, S., Ilic, K., Kellogg, E. A., McCouch, S., Pujar, A., Reiser, L., Rhee, S. Y., Sachs, M. M., Schaeffer, M., Stein, L., Stevens, P., Vincent, L., Ware, D., and Zapata, F. (2005). Plant ontology (po): a controlled vocabulary of plant structures and growth stages. *Comp Funct Genomics*, **6**(7-8), 388–97.

Jensen, L., Saric, J., and Bork, P. (2006). Literature mining for the biologist: from information retrieval to biological discovery. *Nature Reviews Genetics*, **7**, 119–27.

Jiang, J. and Conrath, D. (1997). Semantic similarity based on corpus statistics and lexical taxonomy. In *Proceedings of the International Conference on Research in Computational Linguistics (ROCLING X)*, pages 19–33.

Kaplan, A. (1955). An experimental study of ambiguity and context. *Mechanical Translation*, **2**(2), 39–46.

Keok, L. and Ng, H. (2002). An empirical evaluation of knowledge sources and learning algorithms for word sense disambiguation. In *Proceedings of the 2002 Conference on Empirical Methods in Natural Language Processing (EMNLP, Philadelphia, PA)*, pages 41–48.

Khelif, K., Dieng-Kuntz, R., and Barbry, P. (2007). An ontology-based approach to support text mining and information retrieval in the biological domain. *In Special Issue on Ontologies and their Applications of the Journal of Universal Computer Science (JUCS)*, **13**(12), 1881–1907.

Khelif, K., Gandon, F. L., Corby, O., and Dieng-Kuntz, R. (2008). Using the intension of classes and properties definition in ontologies for word sense disambiguation. In A. Gangemi and J. Euzenat, editors, *EKAW*, volume 5268 of *Lecture Notes in Computer Science*, pages 188–197. Springer.

Kilgarriff, A. and Grefenstette, G. (2003). Introduction to the special issue on the Web as corpus. *Computat. Ling.*, **29**(3), 333–347.

Kilgarriff, A. and Palmer, M. (2000). Introduction to the special issue on senseval. *Computers and the Humanities*, **34**(1–2), 1–13.

Kleinjan, D. A., Seawright, A., Mella, S., Carr, C. B., Tyas, D. A., Simpson, T. I., Mason, J. O., Price, D. J., and Heyningen, V. v. H. (2006). Long-range downstream enhancers are essential for pax6 expression. *Dev Biol*, **299**(2), 563–81.

Klinkenberg, R. and Joachims, T. (2000). Detecting Concept Drift with Support Vector Machines. In *Proceedings of the Seventeenth International Conference on Machine Learning (ICML)*, pages 487–494.

Kučera, H. and Francis, W. (1967). *Computational Analysis of Present-Day American English*. Brown University Press, Providence, RI.

Lame, G. (2004). Using nlp techniques to identify legal ontology components: Concepts and relations. *Artificial Intelligence and Law*, **12**, 379–396.

Leacock, C. and Chodorow, M. (1998). Combining local context and WordNet similarity for word sense identification. In C. Fellbaum, editor, *WordNet: An Electronic Lexical Database*, pages 305–332, Cambridge, MA. The MIT Press.

Leacock, C., Towell, G., and Voorhees, E. (1993). Corpus-based statistical sense resolution. In *Proceedings of the ARPA Workshop on Human Language Technology (Princeton, NJ)*, pages 260–265.

Leacock, C., Chodorow, M., and Miller, G. (1998). Using corpus statistics and WordNet relations for sense identification. *Computat. Ling.*, **24**(1), 147–166.

Lee, Y. K., Ng, H. T., and Chia, T. K. (2004). Supervised word sense disambiguation with support vector machines and multiple knowledge sources. In *Senseval-3: Third International Workshop on the Evaluation of Systems for the Semantic Analysis of Text*, pages 137–140.

Leroy, G. and Rindflesch, T. (2005). Effects of information and machine learning algorithms on word sense disambiguation with small datasets. *Int J Med Inform*, **74**(7-8), 573–85.

Lesk, M. (1986). Automated sense disambiguation using machine-readable dictionaries: How to tell a pine cone from an ice cream cone. In *Proceedings of the 1986 ACM SIGDOC Conference, Toronto, Canada*, pages 24–26.

Lin, D. (1998). An information-theoretic definition of similarity. In *Proc. 15th International Conf. on Machine Learning*, pages 296–304. Morgan Kaufmann, San Francisco, CA.

Liu, H., Johnson, S. B., and Friedman, C. (2002). Automatic resolution of ambiguous terms based on machine learning and conceptual relations in the UMLS. *J Am Med Inform Assoc*, **9**(6), 621–636.

Liu, H., Teller, V., and Friedman, C. (2004). A multi-aspect comparison study of supervised word sense disambiguation. *J Am Med Inform Assoc*, **11**(4), 320–331.

Lord, P. W., Stevens, R. D., Brass, A., and Goble, C. A. (2003). Investigating semantic similarity measures across the Gene Ontology: the relationship between sequence and annotation. *Bioinformatics*, **19**(10), 1275–1283.

Madhu, S. and Lytle, D. W. (1965). A figure of merit technique for the resolution of non-grammatical ambiguity. *Mechanical translation*, **8**(2), 9–13.

Magnini, B. and Cavaglià, G. (2000). Integrating subject field codes into WordNet. In *Proceedings of the 2nd Conference on Language Resources and Evaluation (LREC, Athens, Greece)*, pages 1413–1418.

Manning, C. and Schütze, H. (1999). *Foundations of Statistical Natural Language Processing*. MIT Press, Cambridge, MA.

Masterman, M. (1957). The thesaurus in syntax and semantics. *Mechanical Translation*, **4**(1–2), 35–43.

McCray, A. T., Srinivasan, S., and Browne, A. C. (1994). Lexical methods for managing variation in biomedical terminologies. *Proc Annu Symp Comput Appl Med Care*, pages 235–239.

McCulloch, W. and Pitts, W. (1943). A logical calculus of the ideas immanent in nervous activity. *Bull. Math. Biophys.*, **5**, 115–133.

Mihalcea, R. (2004). Co-training and self-training for word sense disambiguation. In *Proceedings of CoNLL-2004*, pages 33–40. Boston, MA, USA.

Mihalcea, R. (2006). *Word Sense Disambiguation: Algorithms and Applications*, chapter Knowledge-based methods for WSD, pages 107–131. Springer, New York.

Mihalcea, R. and Edmonds, P., editors (2004). *Proceedings of the 3rd InternationalWorkshop on the Evaluation of Systems for the Semantic Analysis of Text (Senseval-3, Barcelona, Spain)*.

Mihalcea, R., Tarau, P., and Figa, E. (2004). Pagerank on semantic networks, with application to word sense disambiguation. In *Proceedings of the 20th International Conference on Computational Linguistics (COLING, Geneva, Switzerland)*, pages 1126–1132.

Miller, G., Leacock, C., and Bunker, R. T. R. (1993). A Semantic Concordance. In *Proceedings of the ARPA Workshop on Human Language Technology*, pages 303–308.

Mooney, R. (1996). Comparative experiments on disambiguating word senses: An illustration of the role of bias in machine learning. In *Proceedings of the 1996 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 82–91.

Morris, J. and Hirst, G. (1991). Lexical cohesion computed by thesaural relations as an indicator of the structure of text. *Comput. Linguist.*, **17**(1), 21–48.

Mueller, H.-M., Kenny, E. E., and Sternberg, P. W. (2004). Textpresso: An ontology-based information retrieval and extraction system for biological literature. *PLoS Biology*, **2**(11), e309.

Mukherjea, S. (2005). Information retrieval and knowledge discovery utilising a biomedical semantic web. *Brief Bioinform*, **6**(3), 252–262.

Murata, M., Utiyama, M., Uchimoto, K., Ma, Q., and Isahara, H. (2001). Japanese word sense disambiguation using the simple Bayes and support vector machine methods. In *Proceedings of the 2nd International Workshop on Evaluating Word Sense Disambiguation Systems (Senseval-2, Toulouse, France)*, pages 135–138.

Navigli, R. (2009). Word sense disambiguation: A survey. *ACM Comput. Surv.*, **41**(2), 1–69.

Navigli, R. and Velardi, P. (2005). Structural semantic interconnections: A knowledge-based approach to word sense disambiguation. *IEEE Trans. Patt. Anal. Mach. Intell.*, **27**(7), 1075–1088.

Navigli, R. and Verlardi, P. (2004). Learning domain ontologies from document warehouses and dedicated websites. *Computational Linguistics*, **30**(2), 151–179.

Navigli, R., Velardi, P., and Gangemi, A. (2003). Ontology learning and its application to automated terminology translation. *IEEE Intelligent Systems*, **18**(1), 22–31.

Nelson, S., Johnston, D., and Humphreys, B. (2001). *Relationships in the organization of knowledge*, chapter Relationships in Medical Subject Headings, pages 171–184. New York: Kluwer Academic Publishers.

Newman, M. E. J. (2004). Coauthorship networks and patterns of scientific collaboration. *Proc Natl Acad Sci U S A*, **101 Suppl 1**, 5200–5.

Ng, H. and Lee, H. (1996). Integrating multiple knowledge sources to disambiguate word senses: An examplar-based approach. In *Proceedings of the 34th Annual Meeting of the Association for Computational Linguistics (Santa Cruz, CA)*, pages 40–47.

Ng, T. (1997). Getting serious about word sense disambiguation. In *Proceedings of the ACL SIGLEX Workshop on Tagging Text with Lexical Semantics: Why, What, and How?*, pages 1–7, Washington D.C.

Ng, V. and Cardie, C. (2003). Weakly supervised natural language learning without redundant views. In *Proceedings of the Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics (HLT/NAACL, Edmonton, Alta., Canada)*, pages 173–180.

Novischi, A., Srikanth, M., and Bennett, A. (2007). Lcc-wsd: system description for english coarse grained all words task at semeval 2007. In *SemEval '07: Proceedings of the 4th International Workshop on Semantic Evaluations*, pages 223–226, Morristown, NJ, USA. Association for Computational Linguistics.

Nowicki, S. (2003). Student vs. search engine: Undergraduates rank results for relevance. *Libraries and the Academy*, **3**, 503–515.

Ogren, P. V., Cohen, K. B., Acquaah-Mensah, G. K., Eberlein, J., and Hunter, L. (2004). The compositional structure of gene ontology terms. *Pac Symp Biocomput*, pages 214–225.

Ogren, P. V., Cohen, K. B., and Hunter, L. (2005). Implications of compositionality in the gene ontology for its curation and usage. *Pac Symp Biocomput*, pages 174–185.

Oliver, H., Diallo, G., de Quincey, E., Kostkova, P., Jawaheer, G., Alexopoulou, D., Habermann, B., Stevens, R., Jupp, S., Khelif, K., Schroeder, M., and Madle, G. (2009). A user-centred evaluation framework for the sealife semantic web browsers. *BMC Bioinformatics*, **10**, S14. special issue dedicated to the SWAT4LS workshop.

Pahikkala, T., Ginter, F., Boberg, J., Järvnen, J., and Salakoski, T. (2005). Contextual weighting for Support Vector Machines in literature mining: an application to gene versus protein name disambiguation. *BMC Bioinformatics*, **6**(157).

Pedersen, T. (1998). *Learning probabilistic models of word sense disambiguation*. Ph.D. thesis, Southern Methodist University, Dallas, TX.

Pedersen, T. and Bruce, R. (1997). Distinguishing word senses in untagged text. In *Proceedings of the Second Conference on Empirical Methods in Natural Language Processing*, pages 197–207, Providence, RI.

Pedersen, T. and Bruce, R. (1998). Knowledge lean word sense disambiguation. In *Proceedings of the Fifteenth National Conference on Artificial Intelligence*, pages 800–805, Madison, WI.

Perez-Iratxeta, C., Pérez, A., Bork, P., and Andrade, M. (2003). Update on xplormed: A web server for exploring scientific literature. *Nucleic Acids Res*, **31**(13), 3866–8.

Pesquita, C., Faria, D., ao, A. F., Lord, P., and Couto, F. (2009). Semantic similarity in biomedical ontologies. *PLoS Comput Biol*, **5**(7), e1000443.

Pianta, E., Bentivogli, L., and Girardi, C. (2002). MultiWordNet: Developing an aligned multilingual database. In *Proceedings of the 1st International Conference on Global WordNet*, pages 21–25, Mysore, India.

Pietra, S. D., Pietra, V. J. D., and Lafferty, J. D. (1997). Inducing features of random fields. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **19**(4), 380–393.

Popescul, A. and Ungar, L. (2000). Automatic labeling of document clusters. http://www.cis.upenn.edu/ popescul/Publications/popescul00labeling.pdf.

Porter, M. F. (1980). An algorithm for suffix stripping. *Program*, **14**(3), 130–137.

Price, D. (1965). Network of scientific papers. *Science*, **149**(3683), 510515.

Procter, P., editor (1978). *Longman Dictionary of Contemporary English*. London:Longman Group.

Purandare, A. and Pedersen, T. (2004). Word sense discrimination by clustering contexts in vector and similarity spaces. In *Proceedings of CoNLL-2004*, pages 41–48. Boston, MA, USA.

Quinlan, J. (1993). *Programs for Machine Learning*. Morgan Kaufmann, San Francisco, CA.

Rada, R., Mili, H., Bicknell, E., and Blettner, M. (1989). Development and application of a metric on semantic nets. *Systems, Man and Cybernetics, IEEE Transactions on*, **19**(1), 17–30.

Rector, A. (2004). Defaults, context, and knowledge: alternatives for owl-indexed knowledge bases. *Pac Symp Biocomput*, pages 226–237.

Rector, A., Rogers, J., and Pole, P. (1996). The galen high level ontology. In J. Brender, J. Christensen, J. Scherrer, and P. McNair, editors, *Proceedings of Medical Informatics Europe '96 (MIE'96); Copenhagen*, pages 174–178. IOS Press.

Reichert, M., Linckels, S., Meinel, C., and Engel, T. (2005). Student's perception of a semantic search engine. In S. Kinshuk and P. Isaías, editors, *Proceedings of the IADIS Cognition and Exploratory Learning in Digital Age (CELDA 2005): 14-16 December 2005; Porto, Portugal*, pages 139–147. IADIS.

Resnik, P. (1995). Using information content to evaluate semantic similarity in a taxonomy. In *IJCAI'95: Proceedings of the 14th international joint conference on Artificial intelligence*, pages 448–453, San Francisco, CA, USA. Morgan Kaufmann Publishers Inc.

Resnik, P. (1999). Semantic Similarity in a Taxonomy: An Information-Based Measure and its Application to Problems of Ambiguity in Natural Language. *Journal of Artificial Intelligence Research*, **11**, 95–130.

Resnik, P. and Yarowsky, D. (1999). Distinguishing systems and distinguishing senses: New evaluation methods for word sense disambiguation. *Natural Language Engineering*, **5**(2), 113–133.

Richardson, R. and Smeaton, A. F. (1995). Using WordNet in a knowledge-based approach to Information Retrieval. Technical Report CA-0395, Dublin, Ireland.

Ringwald, M., Eppig, J. T., Begley, D. A., Corradi, J. P., McCright, I. J., Hayamizu, T. F., Hill, D. P., Kadin, J. A., and Richardson, J. E. (2001). The mouse gene expression database (gxd). *Nucl. Acids Res.*, **29**(1), 98–101.

Rivest, R. (1987). Learning decision lists. *Mach. Learn.*, **2**(3), 229–246.

Rosse, C. and Mejino, J. L. V. (2003). A reference ontology for biomedical informatics: the foundational model of anatomy. *J Biomed Inform*, **36**(6), 478–500.

Roy, A., Kostkova, P., Catchpole, M., and Carson, E. (2006). Web-based provision of information on infectious diseases: a systems study. *Health Informatics J*, **12**(4), 274–92.

Schijvenaars, B., Mons, B., Weeber, M., Schuemie, M., van Mulligen, E., Wain, H., and Kors, J. (2005). Thesaurus-based disambiguation of gene symbols. *BMC Bioinformatics*, **6**(1), 149.

Schlicker, A., Domingues, F., Rahnenfuhrer, J., and Lengauer, T. (2006). A new measure for functional similarity of gene products based on Gene Ontology. *BMC Bioinformatics*, **7**(1), 302.

Schroeder, M., Burger, A., Kostkova, P., Stevens, R., Habermann, B., and Dieng-Kuntz, R. (2006). From a service-based escience infrastructure to a semantic web for the life sciences: The sealife project. In G. Armano, L. Milanesi, and P. Romano, editors, *Proceedings of the Workshop on Network Tools and Applications in Biology, NETTAB, Santa Margherita di Pula, Italy*.

Schuemie, M. J., Kors, J. A., and Mons, B. (2005). Word sense disambiguation in the biomedical domain: an overview. *J Comput Biol*, **12**(5), 554–565.

Schulz, S., Markó, K., and Hahn, U. (2007). Spatial location and its relevance for terminological inferences in bio-ontologies. *BMC Bioinformatics*, **8**(1), 134.

Schütze, H. (1998). Automatic word sense discrimination. *Comput. Linguist.*, **24**(1), 97–123.

Schütze, H. and Pedersen, J. (1995). Information retrieval based on word senses. In *Proceedings of the 4th Annual Symposium on Document Analysis and Information Retrieval: 1995;Las Vegas, NV*, pages 161–175.

Shneiderman, B. (1992). *Designing the user interface: Strategies for effective human-computer interaction*. Addison-Wesley Publishing Company.

Skrzypek, M. S., Arnaud, M. B., Costanzo, M. C., Inglis, D. O., Shah, P., Binkley, G., Miyasato, S. R., and Sherlock, G. (2010). New tools at the candida genome database: biochemical pathways and full-text literature search. *Nucleic Acids Res*, **38**(Database issue), D428–32.

Smith, B., Köhler, J., and Kumar, A. (2004). On the application of formal principles to life science data: a case study in the gene ontology. In *DILS*, pages 79–94.

Smith, B., Ceusters, W., Klagges, B., Kohler, J., Kumar, A., Lomax, J., Mungall, C., Neuhaus, F., Rector, A., and Rosse, C. (2005). Relations in biomedical ontologies. *Genome Biology*, **6**(5), R46.

Soldatova, L. and King, R. (2005). Are the current ontologies in biology good ontologies? *Nat Biotechnol*, **23**(9), 1095–8.

Sowa, J. F. (2000). *Knowledge Representation: Logical, Philosophical, and Computational Foundations*. Brooks Cole Publishing Co.

Spackman, K. (2004). Snomed ct milestones: endorsements are added to already-impressive standards credentials. *Healthc Inform*, **21**(9), 54, 56.

Sprague, J., Bayraktaroglu, L., Clements, D., Conlin, T., Fashena, D., Frazer, K., Haendel, M., Howe, D. G., Mani, P., Ramachandran, S., Schaper, K., Segerdell, E., Song, P., Sprunger, B., Taylor, S., Van Slyke, C. E., and Westereld, M. (2006). The zebrash information network: the zebrash model organism database. *Nucleic Acids Res*, **34**(Database issue), 581585.

Stefanowski, J. and Weiss, D. (2003). Carrot2 and language properties in web search results clustering. In E. Ruiz, J. Segovia, and P. Szczepaniak, editors, *Proceedings of the First International Atlantic Web Intelligence Conference: 5-6 May 2003; Madrid, Spain*, pages 240–249. Springer.

Stein, L., Sternberg, P., Durbin, R., Thierry-Mieg, J., and Spieth, J. (2001). Wormbase: network access to the genome and biology of caenorhabditis elegans. *Nucleic Acids Res*, **29**(1), 82–86.

Su, L. (2003). A comprehensive and systematic model of user evaluation of web search engines: Ii. an evaluation by undergraduates. *J Am Soc Inf Sci*, **54**, 1193–1223.

Sussna, M. (1993). Word sense disambiguation for free-text indexing using a massive semantic network. In *CIKM '93: Proceedings of the second international conference on Information and knowledge management*, pages 67–74, New York, NY, USA. ACM Press.

Thut, C. J., Rountree, R. B., Hwa, M., and Kingsley, D. M. (2001). A large-scale in situ screen provides molecular evidence for the induction of eye anterior segment structures by the developing lens. *Dev Biol*, **231**(1), 63–76.

Tomanek, K., Wermter, J., and Hahn, U. (2007). An approach to text corpus construction which cuts annotation costs and maintains reusability of annotated data. In *Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL)*, pages 486–495, Prague, Czech Republic. Association for Computational Linguistics.

Towell, G. and Voorhees, E. (1998). Disambiguating highly ambiguous words. *Computat. Ling.*, **24**(1), 125–145.

Tsatsaronis, G., Vazirgiannis, M., and Androutsopoulos, I. (2007). Word sense disambiguation with spreading activation networks generated from thesauri. In *Proceedings of the 20th International Joint Conference on Artificial Intelligence (IJCAI, Hyderabad, India)*, pages 1725–1730.

Tsatsaronis, G., Varlamis, I., and Vazirgiannis, M. (2010). Text Relatedness Based on a Word Thesaurus. *Journal of Artificial Intelligence Research*, **37**, 1–39.

Uren, V., Motta, E., Dzbor, M., and Cimiano, P. (2005). Browsing for information by highlighting automatically generated annotations: user study and evaluation. In P. Clark and G. Schreiber, editors, *Proceedings of the Third International Conference on Knowledge Capture: 2-5 October 2005; Banff, Canada.*, pages 75–82. ACM Press.

Uschold, M. (1996). Building ontologies: Towards a unified methodology. In *16th Annual Conf. of the British Computer Society Specialist Group on Expert Systems*, pages 16–18, Cambridge, UK.

Vaughan, L. (2004). New measurements for search engine evaluation proposed and tested. *Information Processing and Management: an International Journal*, **40**, 677–691.

Véronis, J. and Ide, N. (1990). Word sense disambiguation with very large neural networks extracted from machine readable dictionaries. In *Proceedings of the 13th International Conference on Computational Linguistics (COLING, Helsinki, Finland)*, pages 389–394.

W3C Interest Group (2008). W3C Interest Group Note 4 June 2008.

Wächter, T. (2010). Ph.D. thesis, TU Dresden.

Wächter, T., Alexopoulou, D., Dietze, H., Hakenberg, J., and Schroeder, M. (2007). *Anatomy Ontologies for Bioinformatics*, chapter Searching Biomedical Literature with Anatomy Ontologies. Springer.

Weaver, W. (1955). *Machine Translation of Languages*, chapter Translation, pages 15–23. John Wiley & Sons, New York.

Weeber, M., Mork, J. G., and Aronson, A. R. (2001). Developing a Test Collection for Biomedical Word Sense Disambiguation. In *Proc AMIA Symp*, pages 746–750.

Wermter, J., Tomanek, K., and Hahn, U. (2009). High-performance gene name normalization with GeNo. *Bioinformatics*, **25**(6), 815–21.

Whetzel, P. L., Parkinson, H., Causton, H. C., Fan, L., Fostel, J., Fragoso, G., Game, L., Heiskanen, M., Morrison, N., Rocca-Serra, P., Sansone, S.-A., Taylor, C., White, J., and Stoeckert, C. J. (2006). The mged ontology: a resource for semantics-based description of microarray experiments. *Bioinformatics*, **22**(7), 866–873.

Widdows, D., Peters, S., Cederberg, S., Chan, C.-K., Steffen, D., and Buitelaar, P. (2003). Unsupervised Monolingual and Bilingual Word-Sense Disambiguation of Medical Documents using UMLS. In *Proceedings of the ACL Workshop on Natural Language Processing in Biomedicine*, pages 9–16, Sapporo, Japan. Association for Computational Linguistics.

Wilks, Y. (1975). *Formal Semantics of Natural Language*, chapter Preference semantics, pages 329–348. Cambridge, UK: Cambridge University Press.

Wilks, Y., Fass, D., Guo, C.-M., MacDonald, J. E., Plate, T., and Slator, B. A. (1990). *Semantics and the Lexicon*, chapter Providing machine tractable dictionary tools, pages 341–401. Kluwer Academic Publishers.

Winnenburg, R., Wächter, T., Plake, C., Doms, A., and Schroeder, M. (2008). Facts from text: can text mining help to scale-up high-quality manual curation of gene products with ontologies? *Brief Bioinform*, **9**(6), 466–78.

Wren, J., Chang, J., Pustejovsky, J., Adar, E., Garner, H., and Altman, R. (2005). Biomedical term mapping databases. *Nucleic Acids Research*, **33**, D289–93.

Wroe, C., Stevens, R., Goble, C., and Ashburner, M. (2003). A methodology to migrate the gene ontology to a description logic environment using daml+oil. *Pac Symp Biocomput*, page 624636.

Xu, H., Markatou, M., Dimova, R., Liu, H., and Friedman, C. (2006). Machine learning and word sense disambiguation in the biomedical domain: design and evaluation issues. *BMC Bioinformatics*, **7**(1), 334.

Yarowsky, D. (1992). Word sense disambiguation using statistical models of roget's categories trained on large corpora. In *Proceedings of the 14th International Conference on Computational Linguistics (COLING)*, pages 454–460, Nantes, France.

Yarowsky, D. (1993). One sense per collocation. In *HLT '93: Proceedings of the workshop on Human Language Technology*, pages 266–271, Morristown, NJ, USA. Association for Computational Linguistics.

Yarowsky, D. (1995). Unsupervised word sense disambiguation rivaling supervised methods. In *Meeting of the Association for Computational Linguistics*, pages 189–196.

Yarowsky, D. (2000). Hierarchical decision lists for word sense disambiguation. *Comput. Human.*, **34**(1-2), 179–186.

Yeh, A., Hirschman, L., and Morgan, A. (2002). Background and overview for kdd cup 2002 task 1: information extraction from biomedical articles. *SIGKDD Explor. Newsl.*, **4**(2), 87–89.

Yesilada, Y., Bechhofer, S., and Horan, B. (2008). *Advances in Semantic Media Adaptation and Personalization.*, chapter Dynamic Linking of Web Resources: Customisation and Personalisation., pages 1–24. Springer.

Yu, H., Hripcsak, G., and Friedman, C. (2002). Mapping Abbreviations to Full Forms in Biomedical Articles. *Journal of the American Medical Informatics Association*, **9**(3), 262–272.

Yu, H., Kim, W., Hatzivassiloglou, V., and Wilbur, W. J. (2007). Using MEDLINE as a knowledge source for disambiguating abbreviations and acronyms in full-text biomedical journal articles. *J. of Biomedical Informatics*, **40**(2), 150–159.

Zavitsanos, E., Paliouras, G., Vouros, G., and Petridis, S. (2007). Discovering Subsumption Hierarchies of Ontology Concepts from Text Corpora. In *Web Intelligence*, pages 402–408. IEEE Computer Society.

Zeeberg, B. R., Feng, W., Wang, G., Wang, M. D., Fojo, A. T., Sunshine, M., Narasimhan, S., Kane, D. W., Reinhold, W. C., Lababidi, S., Bussey, K. J., Riss, J., Barrett, J. C., and Weinstein, J. N. (2003). Gominer: a resource for biological interpretation of genomic and proteomic data. *Genome Biol*, **4**(4), R28.

Zipf, G. (1949). *Human Behaviour and the Principle of Least Effort: An introduction to human ecology*. Addison-Wesley. Reprinted by New York: Hafner, 1972.